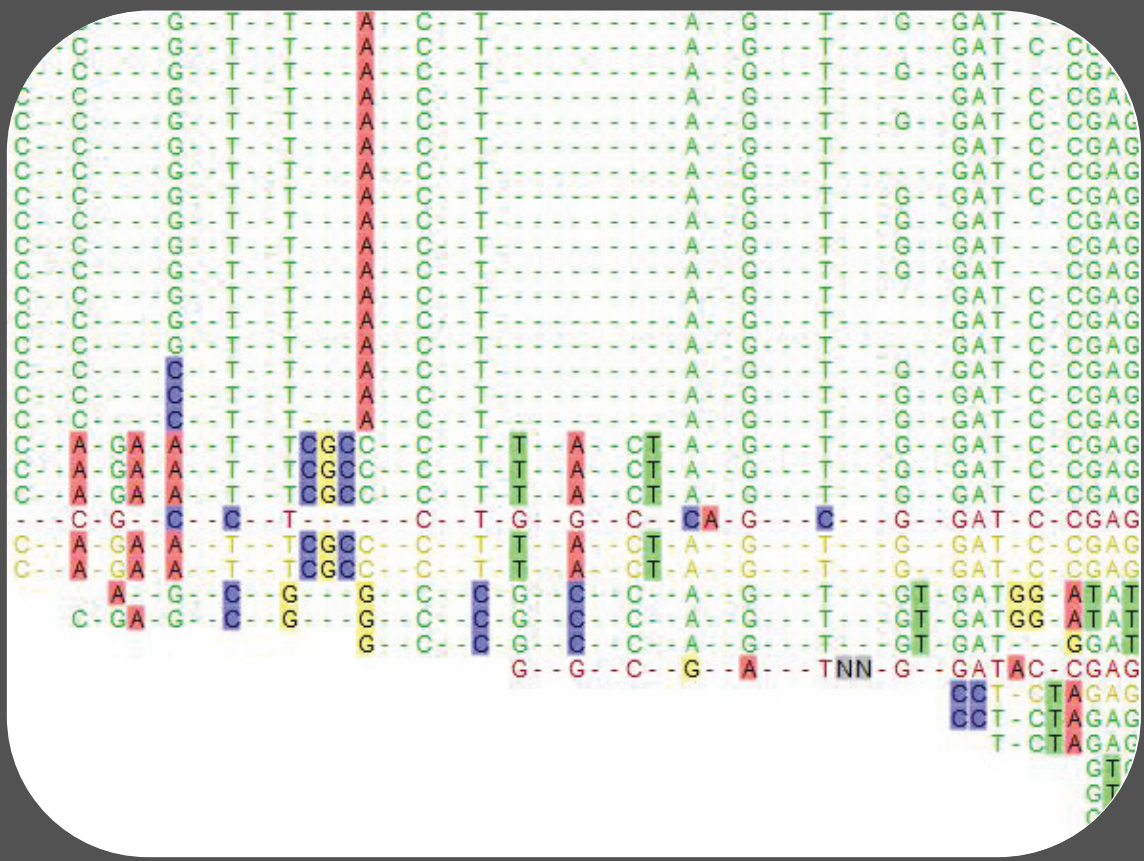
On the road of sequencing the genomes: Past, Present and Future

Alcino Orfeu de Leão e Flores



Dissertation presented to obtain the Ph.D degree in Molecular Genetics

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras,
Dezembro, 2015

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
ANTÓNIO XAVIER /UNL

Knowledge Creation

On the road of sequencing the genomes: Past, Present and Future

Alcino Orfeu de Leão e Flores

Dissertation presented to obtain the Ph.D degree in Molecular Genetics
Instituto de Tecnologia Química e Biológica António Xavier
Universidade Nova de Lisboa

Oeiras, Dezembro, 2015

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
ANTÓNIO XAVIER /UNL

Knowledge Creation

Supervisor:

Professora Doutora Claudina Rodrigues-Pousada
Professora Catedrática Convidada
Head of the Genomics and Stress Laboratory
Instituto de Tecnologia Química e Biológica - António Xavier, UNL

Jury Members:

President: Professor Doutor Cláudio Soares, Director,
Instituto de Tecnologia Química e Biológica - António Xavier, UNL

Professora Doutora Helena Santos,
Instituto de Tecnologia Química e Biológica - António Xavier, UNL

Professora Doutora Cecília Leão, Director
Escola de Ciências da Saúde - Universidade do Minho

Doutora Inês Pereira, Vice-Director
Instituto de Tecnologia Química e Biológica - António Xavier, UNL

Professor Doutor António Alfredo Coelho Jacinto, Coordinator
Centro de Estudos de Doenças Crónicas, Nova Medical School - UNL

Dedicado a todos os membros das famílias Leão e Flores.

Dedicated to all Leão and Flores family members.

Acknowledgments

I acknowledge the strong spirit and inexhaustible scientific drive and professionalism of Prof. Claudina Rodrigues-Pousada, for the supervision of this thesis. Her example is a true inspiration for any scientific career.

I am also in the true debt to all STAB VIDA staff members, who have joined me in the most beautiful, but demanding dream. For their belief and dedication, I kindly thank the team: Adrian Posado, Alvaro Miro, Carla Clemente, Carlos Bernardes, Daniela Leão, Fabio Carlos, Filipa Patriarca, Gonçalo Doria, Hugo Pereira, Inês Figueiredo, Juan Coronado, Leonor Soares, Lídia Cubero, Liliana Castro, Magdalena Lewicka, Marcia Matos, Nanci Lopes, Nuno Granadeiro, Paulo Almeida, Patrícia Topete, Pedro Penedo, Pedro Pinto, Rui Crespo, Sofia Goes, Sofia Marcos, Tatiana Flores and Vitaliy Sobchuk.

A special and warm acknowledgment is due to ITQB, for being a driver of excellence and a pillar for knowledge creation. I thank the ITQB for hosting and supporting great part of the experimental work, performed in the laboratory of "Stress and Genomics", where I met the most dedicated and brilliant young scientists: Gabriela Silva, Rute Marques, Manuela Broco, Cátia Santos, Fabio Silva and Catarina Pimentel.

For similar reasons I want to thank the Instituto Gulbenkian de Ciência, for hosting the first years of the experimental work performed.

The excellent bioinformatics work performed, fundamental for delivering all the important findings presented in this work, makes me in debt to Jeronimo Ruiz team of Belo Horizonte and Paulo Almeida of STAB VIDA.

Pedro Lança has my great recognition for helping editing this thesis.

Finally, I thank the former "JNICT" (Junta Nacional de Investigação Científica) for my PhD grant (many years ago) as well as AdI (Agência de Inovação) and STAB VIDA for their financial support to the Gigasnoma project.

Summary

Desulfovibrio gigas is a model organism of sulfate-reducing bacteria of which energy metabolism and stress response have been extensively studied. The complete genomic context of this organism was however not yet available. This thesis is about the sequencing of the *D. gigas* genome and provides insights into the integrated network of energy conserving complexes and structures present in this bacterium. The work presented describes the technological processes used to sequence the genome. Chapter I is an overview of its most recent advances in bacterial genome sequencing, since the first genome published of *Haemophilus influenzae* till present; Chapter II is mostly an overview of the 4 different technologies that were necessary to finalize this genome sequencing project: shotgun sequencing through Sanger method and massive DNA sequencing from 3 different next generation sequencing (NGS) platforms: 454 (Roche); Genome Analyzer and Hiseq2000 (Illumina) and Ion Torrent (Life Technologies). Chapter III is an exercise of bioinformatics for comparative analysis of all different sources of raw data of DNA reads, and for assembling the DNA bacterial chromosome and plasmid. Chapter IV is dedicated to the results of assembling and annotating the genome of *Desulfovibrio gigas*, and some important findings and location of genes coding for metallo-proteins; And finally, chapter V is dedicated to the trends in genomics and to the importance of translational research. It should be emphasized that this thesis, besides delivering the main findings on *D. gigas* genome, intends to propose a model for fostering translational research in Portuguese research Institutes and Universities, taking advantage of the author's accumulated experience on this matter, being this work presented as an example of a successful industry-academia partnership.

Sumário

Desulfovibrio gigas é um organismo modelo de bactérias redutoras de sulfato, de quem o matabolismo energético e resposta ao stress têm sido extensivamente estudados. A genómica completa deste organismo foi no entanto apenas disponibilizado no ano de 2014. A sequenciação do genoma de *D. gigas* forneceu importante informação sobre a rede integrada de complexos de conservação de energia e de estruturas presentes nesta bactéria. O trabalho apresentado nesta tese descreve os processos tecnológicos utilizados para sequenciar o genoma da bactéria redutora de sulfato *Desulfovibrio gigas*. O Capítulo I é uma visão geral dos mais recentes avanços na sequenciação do genoma bacteriano, desde o primeiro genoma publicado (*Haemophilus influenzae*) até ao presente; O Capítulo II apresenta uma visão geral das quatro tecnologias diferentes que foram necessárias para finalizar o projeto de sequenciação do genoma: Sanger shotgun, e sequenciação massiva de DNA realizadas em 3 diferentes plataformas de sequenciação de nova geração (NGS): 454 (Roche); Genome Analyzer e Hiseq2000 (Illumina) e Ion Torrent (Life Technologies). O Capítulo III é um exercício de bioinformática com vista à análise comparativa de todas as diferentes fontes de dados brutos de sequências de DNA, e de forma a testar estes mesmos dados na montagem do cromossoma bacteriano e do plasmídeo de *D. gigas*. O Capítulo IV é dedicado à apresentação dos resultados de montagem e anotação do genoma, bem como da localização de genes que codificam para importantes metalo-proteínas. E, finalmente, o capítulo V é dedicado às tendências em genómica e à importância da investigação de translação. Deve-se ressaltar que essa tese, além de introduzir as principais descobertas sobre o genoma de *D. gigas*, tem a intenção de propor um modelo de fomento à investigação translacional em Institutos de investigação Portugueses e em Universidades. Tal é possível recorrendo à experiência acumulada do autor, sendo o presente trabalho é apresentado como um exemplo bem sucedido de uma parceria entre empresa e Universidade.

Index

Figure Index

Table Index

On the road of sequencing the genomes: Past, Present and Future

Preamble

Genome sequencing started as an important scientific challenge and is now becoming a widespread routine in many laboratories, due to its utmost importance in solving research hypothesis and to the availability of massive sequencing technologies.

The number of bacterial genomes that have been sequenced is now (April 2014) a total of 21,820, after the first published genome in 1995 of the bacteria *Haemophilus influenzae*, and there are currently 11,746 bacterial genome sequencing projects in progress.

Small genome sequencing can now be performed, from DNA to raw data, in 2 or 3 weeks, if using the utmost high throughput technologies. Analysing the BIG DATA generated by these technologies requires complex bioinformatics algorithms, that are more and more accurate and fast on assembling the contigs and scaffolds, predicting the annotation, and proposing the encoded metabolic pathways.

This thesis is about the sequencing of the bacterial genome of *Desulfovibrio gigas*, its bioinfomatics analysis and annotation. Using it as an example, it tries to provide the state of art on the overall thematic of small genomes sequencing, some of its numbers and statistics, importance and trends for the future.

Issues of economic importance and relevance of the activity of small genomes sequencing are also presented, introducing the biotech start-up company STAB VIDA ltd as a real example of translational research.

"*All things are difficult before they become easy*"
(Thomas Fuller)

Chapter I – Introducing the bacterial genomes sequencing, the genus ***Desulfovibrio*** and the bacteria ***Desulfovibrio gigas***

Preface

Chapter I describes the evolution seen in the recent years in the field of bacterial genomes research and, in particular, of the *Desulfovibrio* genus. This knowledge has been booming since the year of 2007, and its impact on different fields and applications is very high, has mentioned in studies referred in the next sections

"*An expert is a person who has made all the mistakes that can be made in a very narrow field*"

Niels Bohr

I.1 – State of the art of bacterial genome sequencing

At the moment the Genomes Online Database lists a total of 64,817 studied genomes when considering all the phylogenetic groups, from which 6,410 are defined as complete and published, 820 listed as complete and 17,151 listed as permanent drafts (genomes that are sequenced but not completely closed). Out of that, 30,530 are defined as incomplete, 7,156 are being proposed, 1,044 are targeted, and 1,706 are defined as abandoned *(source: GOLD database)*.

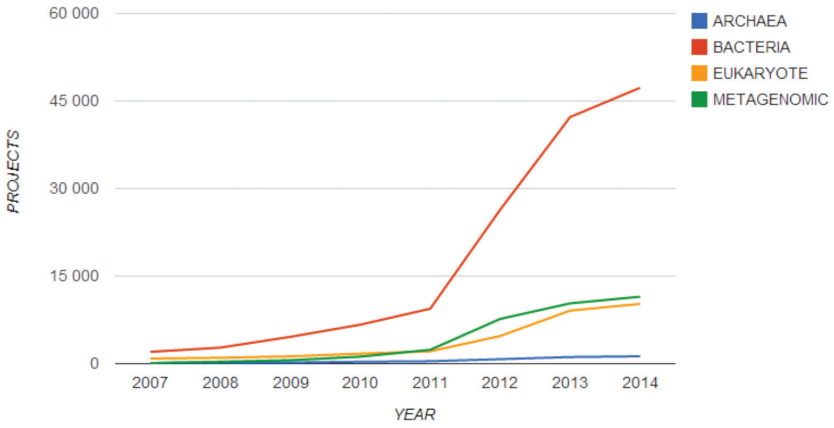


Figure 1 – *Whole genome sequencing projects developed, organized by Phylogenetic Group.*

When considering only bacterial genomes there is a total of 47,256 genomes, from which 2,826 are defined as complete and published, 733 listed as complete and 15,772 listed as permanent drafts (genomes that are sequenced but not completely closed), 19,127 are defined as incomplete. Out of the above total 6,970 are being proposed, 629 are targeted, and 1,199 are defined as abandoned.

Table 1 – Genome sequencing project status of the overall genomes from all the phylogenetic and of the bacterial genomes, specifically.

	COMPLETE AND PUBLISHED	COMPLETE	PERMANENT DRAFTS	PROPOSED	TARGETED	INCOMPLETE	ABANDONED	TOTAL
TOTAL GENOMES	6,410	820	17,151	7,156	1044	30,530	1,706	64,817
BACTERIAL GENOMES	2,826	733	15,772	6,970	629	19,127	1,199	47,256

The most active genomic centres in the world performing bacterial sequencing and the phylogenetic distribution of the target organisms are shown in Fig 2 and Fig 3. This distribution is a direct consequence of the most important phenotypic importance of the selected organisms, as can be seen in Fig.4:

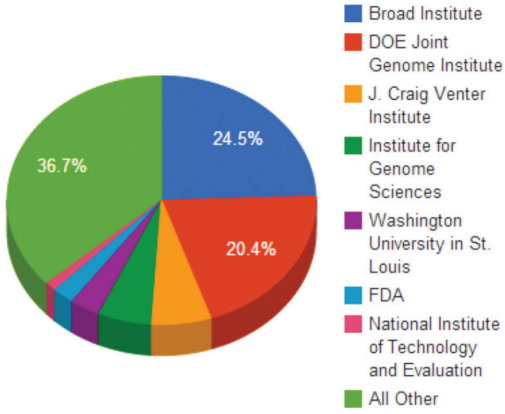


Figure 2 - *Major Sequencing Centres for Archaeal and Bacterial Genomes in 2014, within 39,557 projects (GOLD database).*

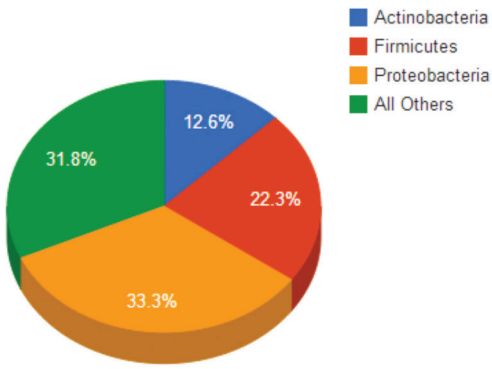


Figure 3 - *Phylogenetic distribution of Bacterial Genome Projects in 2014 from a total of 59,235 projects (GOLD database).*

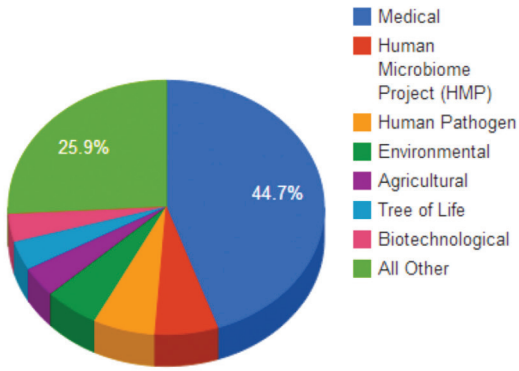


Figure 4 - *Thematic Relevance of Bacterial Genome Projects in 2014 from a total of 39,152 projects (GOLD database).*

The state-of-art of genome sequencing in Portugal is summarized in Table 2. The first *de novo* sequenced genome with the status of complete was the *Desulfovibrio gigas*, deposited in NCBI (CP006585 and its plasmid CP006586) in the year of 2013, and published in 2014 in a open access journal (Morais-Silva *et al* 2014).

Table 2 – The bacterial genomes in Portugal, complete and permanent drafts of species or strains.

GOLDSTAMP	ORGANISM	DE NOVO	SIZE	CONTACT	PROJECT STATUS	COMPLETION DATE	ORFs
Gc0048046	*Chlamydia trachomatis* L2/434/Bu(f) (Re-seq)	Yes	1046 Kb	João P. Gomes	Complete and Published (Borges *et al*, 2013)	2013-06-21	968
Gc0057375	*Desulfovibrio gigas* DSM 1382	Yes	3796 Kb	Rodrigues-Pousada C	Complete and Published (Morais-Silva *et al*, 2013)	2013-09-18	3441
Gi17720	*Pseudomonas sp.* M47T1	No	6311 Kb	Morais, Paula	Permanent Draft	2014-01-08	5753
Gi08982	*Pseudomonas aeruginosa* 138244	No	6529 Kb	Santos PM	Permanent Draft	2014-01-08	6202
Gi0893	*Pseudomonas aeruginosa* 152504	No	6638 Kb	Santos PM	Permanent Draft	2014-01-08	6148
Gi08919	*Pseudomonas sp.* M1	No	6982 Kb	Santos, Pedro	Permanent Draft	2014-01-23	6163
Go0048047	*Chlamydia trachomatis* L2/434/Bu(i) (Re-seq)	Yes	1046 kb	João P. Gomes	Complete and Published (Borges *et al*, 2013)	2014-04-08	968

I.2 - State-of-art of the importance and the applications of bacterial genome sequencing in life sciences in general.

In this section we highlight the importance of bacterial genome sequencing in such diverse areas as archeology research, clinical research, clinical practice, metagenomes for bioremediation and many others. A few illustrative examples of specific bacterial genomes are provided as proof of its contribution to solving important scientific questions.

I.2.1 – Genomic archeology

The change of the human lifestyle into urban environments, with antibiotics and advanced sanitation, represented a fundamental change in the relationship with microbes, and even if we have benefited from that change, it appears the risk for allergies and other inflammatory diseases is increasing. The best way to understand this situation is by studying the ancestral state of the human microbiome (Levy, 2013).

Examples through the detailed analysis of the tRNA structure of the genomes of *Escherichia coli strains* K12, CFT073, and O157:H7, *Shigella flexneri* 2a 301, and *Salmonella typhimurium* LT2, a genomic glance at the problem of tRNA evolution as a repetitive process, and the relationship of this mechanism to genome evolution and codon usage, was done (Withers *et al*, 2006) the authors show that, on average, 0.64 tRNA insertions/duplications occur every milion years per genome per lineage, while deletions occur at the slower rate of 0.30 per million years, per genome per lineage.

I.2.2 - Clinical medicine.

The latest generation of benchtop DNA sequencing platforms can provide an accurate whole-genome sequence (WGS) for a broad range of bacteria in less than a day. There is increasing evidence that use of these techniques could enhance the provision of diagnostic and public health microbiologic analyses and support efforts to understand and contain the spread of microbial pathogens, including multidrug-resistant organisms (Reuter *et al*, 2013).

Example 1: The rapid next-generation technologies facilitated prospective whole genome characterization in the early stages of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak occurred in 2011, helping to make informed decisions about treatment, prevention, and source tracking (Mellman *et al*, 2011).

Example 2: The sequencing of two isolates of *Pseudomonas aeruginosa* obtained from patients with pneumonia, named 138244 and 152504, representative of allelic sequence types ST175 and ST560 respectively, were done in Portugal. Several unique genes encoding hypothetical proteins were found in the isolates, with particular relevance to genes involved in antibiotic-resistance (e.g. puromicin N-acetyltransferase-like gene) (Soares-Castro, 2011).

I.2.3 Metagenomics for Bioremediation

The study of metagenomes obtained directly from the environment is designated as environmental genomics or ecogenomics, where total DNA sequencing from the environmental sample is crucial. We refer, as example, the deep-sea hydrocarbon plumes. Due to crude oil spills, it causes drastic changes in the microbial communities, nevertheless some of those communities

contribute to the bioremediation of the spill. By applying metagenomic approaches to understand the enzymatic processes and community dynamics of oil-degrading bacteria, valuable insights that inform about the ongoing bioremediation efforts can be gained (Pham and Anonye, 2014).

Phytoremediation applications are another example of metagenomics importance. Many omics analyses data are now combined with high-throughput isolation and screening of microbial characteristics to assist in determining the potential activity of microbes that are perhaps not naturally dominant, but that should be targeted in phytoremediation treatments. The characterization of potential microorganisms that have yet to be cultured, in contaminant remediation or plant growth promotion, arises as a promising strategy (Bell et al, 2014).

I.2.4 Metagenomics for Microbiomes (Community genomics)

The multi 'omics' approach is a powerful tool for understanding the functional symbiotic interplay of human eukaryotic and prokaryotic cells and dynamics of molecular modifications of this multi-cellular system in different environmental conditions (Shenderov and Midtvedt, 2014).

An example is the identification and validation of 60,000 type-2-diabetes-associated markers and the establishment of the concept of metagenomic linkage group, enabling taxonomic species-level analyses. MGWAS analysis showed that patients with type 2 diabetes were characterized by a moderate degree of gut microbial dysbiosis, a decrease in the abundance of some universal butyrate-producing bacteria and an increase in various opportunistic pathogens, as well as an enrichment of other microbial functions conferring sulphate reduction and oxidative stress resistance (Qin *et al*, 2012).

I.3 - The genus *Desulfovibrio*: state-of-art of the sequenced genomes of this genus.

Non-photosynthetic eukaryotes are, with a few exceptions, restricted to organic carbon as energy source and oxygen as electron acceptor. On the other hand, prokaryotes explore almost every possible energy source and oxidant available, being also able to sustain life in extreme environments. One of the metabolic strategies adopted is anaerobic respiration, in which organic or inorganic compounds are used as terminal electron acceptor for a respiratory electron transfer chain. The most common oxidants respired by microorganisms are carbon dioxide, sulphate, sulphur, nitrate and iron oxides (Matias *et al*, 2005).

Sulfate-reducing bacteria (SRB), are chemolithotrophic bacteria that use sulfate as terminal electron acceptor, and constitute a unique physiological group of microorganisms that couple anaerobic electron transport to ATP synthesis. The first report of SRB, the *Spirillum desulfuricans* which was later renamed as *Desulfovibrio desulfuricans* was made by Beijerinck, from a Dutch city canal in Delft in 1895. Beijerinck highlighted that microbial formation of hydrogen sulfide has great importance and interest from both scientific and applied points of view. Sewage contamination caused evolution of large amounts of hydrogen sulfide from city canals in summertime. He managed to enrich the "sulfidferment" and to obtain isolated colonies in agar, which were distinct due to their surroundings, constituted by black iron sulfide precipitate. The motile, curved rod morphology that he reported for *Spirillum desulfuricans* were isolated and characterized as the first *Desulfovibrio* species. In 1903 the isolation in pure culture was achieved by van Delded and in 1930 a comprehensive study of its physiology and metabolism was made by Baars (Roy and Trudinger, 1970; Voordouw, 1995; Barton and Fauque, 2009).

The dissimilatory sulphate bacteria are widely distributed in sea water, marine muds, fresh water, soil and oil-bearing environments. They play a major role in anaerobic corrosion processes and other processes of economic importance, and these relevant activities displayed by SRB are a consequence of the unique electron transport components, or the production of high levels of H_2S. The capability of SRB to utilize hydrocarbons in pure cultures and consortia has resulted in using these bacteria for bioremediation of BTEX (benzene, toluene, ethylbenzene and xylene) compounds in contaminated soils. Specific strains of SRB are capable of reducing 3-chlorobenzoate, chloroethenes, or nitroaromatic compounds and this has resulted in proposals to use SRB for bioremediation of environments containing trinitrotoluene and polychloroethenes (Roy and Trudinger, 1970; Barton and Fauque, 2009).

The most easily and rapidly cultured sulfate reducers are the SRB members of the genus *Desulfovibrio*, as such this genus became the most extensive biochemical and physiological model for research. Dissimilatory sulphate reduction in *Desulfovibrio* species is linked to electron transport-coupled phosphorylation because substrate level phosphorylation is inadequate to support their growth. The SRB belonging to the genus of *Desulfovibrio* possess a number of unique biochemical and physiological characteristics such as the requirement of ATP to reduce sulphate, the cytoplasmic localization of two key enzymes [adenylylsulfate (APS) reductase and bisulfite reductase] involved in the pathway of respiratory sulfate reduction, the periplasmic localization of some hydrogenases and the abundance of multihemic c-type cytochromes (Barton and Fauque, 2009).

The investigation of the mechanism of dissimilatory sulfate reduction has been undertaken mostly with *Desulfovibrio* species. Four cytoplasmic enzymes are sufficient for reduction of

sulfate to sulfide in an eight electron reduction process. From a biochemical point of view, SRB are important bacteria because they contain a diversified and complex electron carrier system. A characteristic feature of the sulfate reduction electron transfer pathway is the involvement of multiheme c-type cytochromes and iron–sulfur proteins of low redox potentials (Barton and Fauque, 2009). According to the Genomes Online Database (GOLD) there are currently 14 species of *Desulfovibrio* that are completely sequenced and published among 77 other *Desulfovibrio* species, with the remaining being in an incomplete, or permanent draft stage (please refer to table 3.).

Figure 5 – *Representation of the geographical origin of 77 Desulfovibrio species sequenced, according to the GOLD database. It should be noted that not all of these species are completely sequence (GOLD database - www.genomesonline.org).*

Table 3 - The ***Desulfovibrio*** genus completely sequenced and published genomes

GOLDSTAMP	ORGANISM	SIZE	CONTACT	PROJECT STATUS
Gc00184	*Desulfovibrio vulgaris* Hildenborough	3773 Kb 3642 ORFs	Heidelberg JF	Complete and Published 2014
Gc01109	*Desulfovibrio salexigens* DSM 2638	4290 Kb 3937ORFs	Hazen Terry C	Complete and Published 2013
Gc0000104	*Desulfovibrio africanus* Walvis Bay	4189 Kb 3850 ORFs	Brown Steven	Complete and Published 2013
Gc0057375	*Desulfovibrio gigas* DSM 1382	3796 Kb 3441 ORFs	Rodrigues - Pousada, C	Complete and Published 2014
GC00482	*Desulfovibrio vulgaris* DP4	3661 Kb 3199 ORFs	Walker Christopher B	Complete and Published 2014
Gc01032	*Desulfovibrio magneticus* RS-1	5316 Kb 4760 ORFs	Fujita N	Complete and Published 2013
Gc0000103	*Desulfovibrio sp.* ND132	3859 Kb 3534 ORFs	Brown Steven	Complete and Published 2013
Gc00864	*ThermoDesulfovibrio yellowstonii* DSM 11347	2004 Kb 2083 ORFs	Sutton G	Complete and Published 2014
Gc00315	*Desulfovibrio alaskensis* G20	3730 Kb 3874 ORFs	Wall Judy	Complete and Published 2014
Gc00916	*Desulfovibrio vulgaris* Miyazaki F	4040 Kb 3281 ORFs	Hazen, Terry C	Complete and Published 2013
Gc02510	*Desulfovibrio piezophilus* C1TLV30	3646 Kb 3431 ORFs	Alain Dolla	Complete and Published 2013
Gc00931	*Desulfovibrio desulfuricans* ATCC 27774	2873 Kb 2443 ORFs	Hazen, Terry C	Complete and Published 2013
Gc01651	*Desulfovibrio aespoeensis* Aspo-2	3629 Kb 3405 ORFs	Hazen, Terry C	Complete and Published 2014
Gc01643	*Desulfovibrio vulgaris* RCH1	3734 Kb 3338ORFs	Hazen, Terry C	Complete and Published 2014

I.4 - The bacteria *Desulfovibrio gigas*, and the reasons for deciding to sequence its genome.

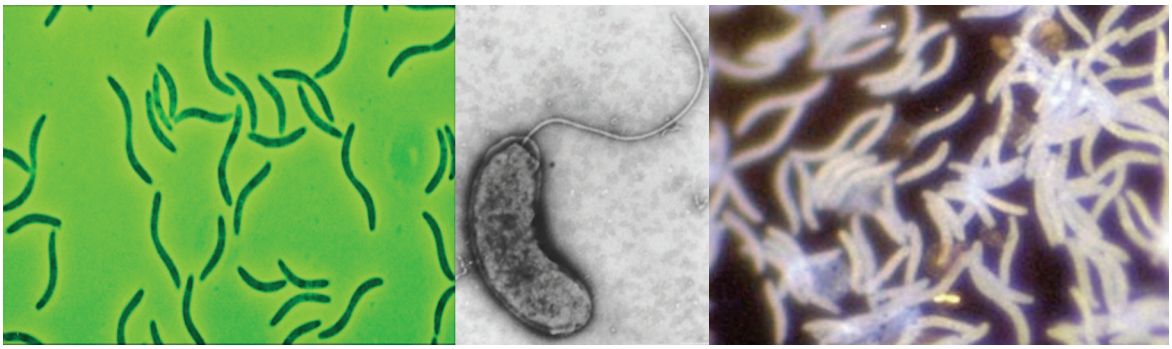


Figure 6 – *The sulfate-reducing bacterium Desulfovibrio gigas. The left and center images were provided by the ITQB and the picture in the right side taken from Water Services LtD.*

Desulfovibrio gigas is remarkable among the sulphate reducing vibrios in several respects: morphologically it is different from other *Desulfovibrios* by its size and its resemblance to a spirillum. In view of its biochemistry, it is one of the best explored anaerobic, nonphotosynthetic heterotrophes with an electron transport phosphorylation system. Furthermore, its species is represented by a single isolate only unlike other *Desulfovibrios* existing in pure culture. Its enrichment and isolation allowed gathering the knowledge about the distribution of this unusual organism (Schoberth, 1993).

It is considered an excellent biological model for investigation of the function and importance of hydrogenases in energy metabolism, since its genome encodes only two hydrogenases, the HynAB and Ech enzymes. Moreover, since each hydrogenase is located in a different cell compartment, *D. gigas* is also an excellent model to study the importance of hydrogen cycling in energy conservation. The *D. gigas* periplasmic HynAB enzyme is one of the most extensively studied enzymes of the [NiFe] type and was the first [NiFe] hydrogenase to have its crystal structure solved (Volbeda *et al*, 1995). The *D. gigas* cytoplasmic Ech hydrogenase belongs to the subgroup of multisubunit membrane-bound energy-conserving [NiFe] hydrogenases (Rodrigues et al, 2003; Morais-Silva *et al*, 2013).

The oxygen-sensing mechanisms of this anaerobic bacterium, indicate a possible transcriptional mechanism to sense and respond to potential stress agents. Previous work using *in vivo* NMR has shown that non-growing cells of *Desulfovibrio gigas* can utilize oxygen as electron acceptor for the reducing power generated during glycogen catabolism, while developing high intracellular levels of ATP. Subsequently, a novel oxygen-utilizing pathway, linking NADH oxidation to oxygen reduction, was elucidated in this organism. This electron-transfer chain comprises two unique proteins, an NADH-rubredoxin oxidoreductase and a new terminal oxygen reductase; it also involves rubredoxin, a well-studied protein to which no physiological function had been previously described. These observations demonstrate the unexpected capacity of this 'strict anaerobe' to profit from the presence of oxygen (Silva et al, 1999; Fareleira *et al,* 2003).

The project of sequencing *D. gigas* genome started officially in 2006, by involving two Portuguese laboratories, both located at Oeiras: ITQB and STAB VIDA, Lda (a biotech start-up). Since 2003 that this consortium had been applying to get funding from the Portuguese "Agência de Inovação", but this agency only considered it as an interesting project for a grant after our two laboratories consortium's third attempt, in the year of 2006.

Years before, Rodrigues-Pousada's laboratory, where I was a PhD student, had delivered the genetic sequence of the very interesting protein MOP – molybdenum aldehyde oxido-reductase, contributing to the knowledge of this organism which, by that time, had more than 300 publications describing its iron-sulphur and other metalloproteinase, with a great portion of those being authored by ITQB researchers. The same happened with *D. gigas* deposited DNA sequences by our research group at Rodrigues-Pousada's lab . In fact, Oeiras was (still is?) the "capital" of *D. gigas*, and sequencing its genome was just a matter of time.

Figure 7 - *Representation of the origin of 14 Desulfovibrio species that have their genome completely sequenced and published, including the Desulfovibrio gigas (GOLD database).*

Chapter II – From Sanger Sequencing to NGS sequencing: the report of how we changed the technological sequencing approach, at the same time that genetics was experiencing a revolution in sequencing methods.

Preface

In the following sections, work that occurred between the years of 2006 and 2013 is presented. The experimental work took longer than expected because, by an unexpected coincidence, the Gigasnoma project progressed in parallel with the NGS revolution that occurred in genetics, from 2005 till now. By 2006 we had no idea of what was coming in the new technologies, and that we would ever need increasingly higher sets of sequence data to complete the project. Here we present how we obtained the four different raw data sets needed to fully close and complete the genome.

Acknowledgements: the work presented in this chapter has, besides the author, the participation of Marcia Matos and Carla Clemente, from STABVIDA's lab, and Morais-Silva and Carla Santos, from Rodrigues-Pousada's lab. Moreover, for the assembling bioinformatics exercise, the author acknowledges the valuable work performed by Jeronimo Ruiz at Fio Cruz's laboratory. Raw data from 454 and Illumina platforms was obtained from subcontracting Biocant (PT), Keygene (NL) and Washindton University (USA).

"*Success consists of going from failure to failure without loss of enthusiasm*"

Winston Churchill

II.1 - The 1st trial for sequencing *D. gigas* genome using shotgun Sanger protocol.

II.1.1 Preface

Our initiative of sequencing the genome of *Desulfovibrio gigas* started in 2006. The idea, then, was to use the same method as Craig Venter did for completing the 1.8Mb genome sequence of the free living organism, *Haemophilus influenza*. Eleven years before, in 1995, this very original work proved that the entire microbial chromosomes could be sequenced rapidly and accurately by applying a shotgun sequencing strategy to whole genomes. In this approach, the DNA is broken up randomly into numerous small segments, such that a single random DNA fragment library is prepared. Then, the ends of a sufficient number of randomly selected fragments are sequenced using the chain termination method, along with capillary electrophoresis, to obtain reads, and assembled to produce a complete genome. Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing. Computer programs then use the overlapping ends of different reads to assemble them into a continuous sequence (Fleischman *et al*, 1995).

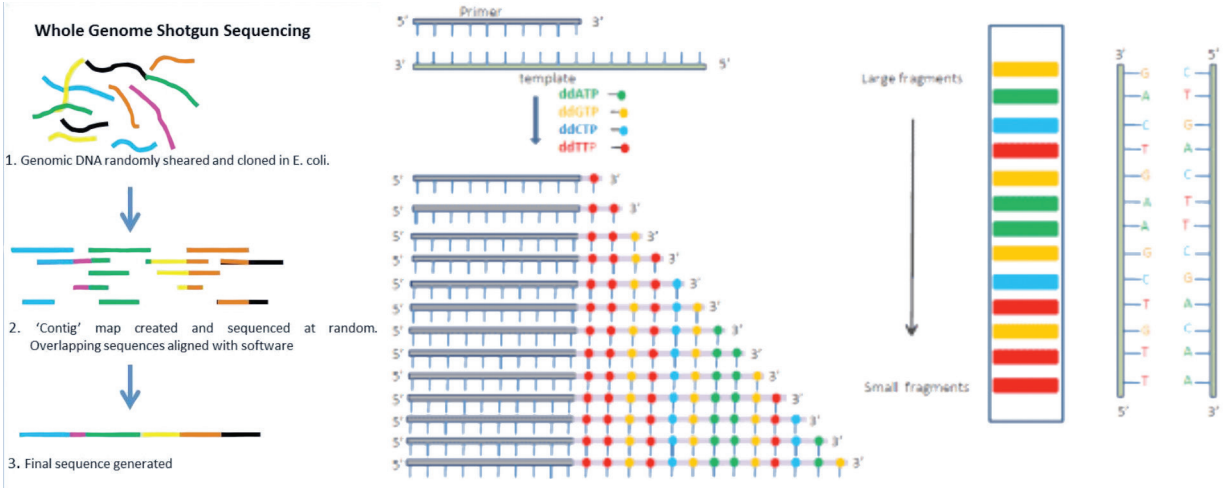


Figure 8 - The innovative method developed in 1995 that become known as "Shot-gun Sequencing.
(http://www.oxbridgebiotech.com/ ; http://esciencecentral.org)

Considered as a breakthrough, in 1995 J. Craig Venter from the Institute for Genomic Research (TIGR) and Nobel laureate Hamilton Smith of Johns Hopkins University, sequenced the 1.8 Mb bacterium *Haemophilus influenzae* with new computational methods developed at TIGR's facility in Gaithersburg, Maryland. Previously, such a shotgunning approach would have failed because the software did not exist to assemble the massive amount of information accurately. The software, developed by TIGR, called the TIGR Assembler was up to the task, reassembling the approximately 24,000 DNA fragments into the whole genome. The TIGR Assembler software required approximately 30 hours of central processing unit time, testifying to the enormous complexity of the computation (Sutton *et al*, 1995)

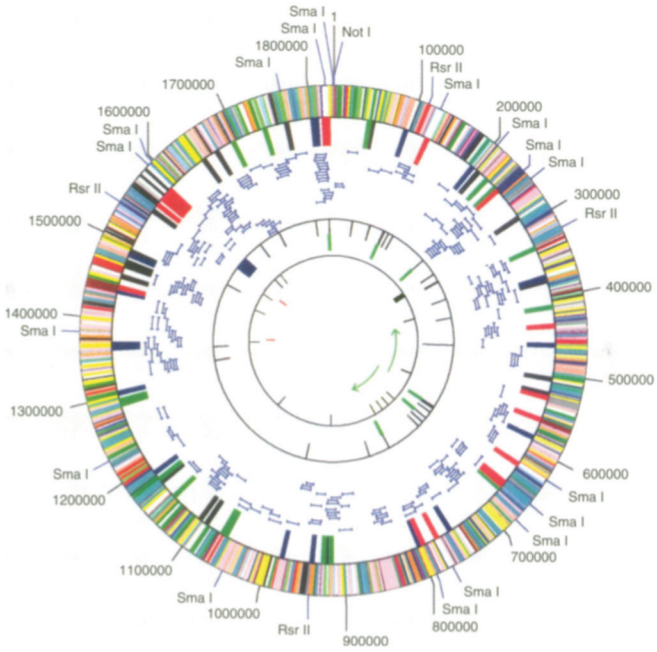


Figure 9 - *Circular representation of the H. influenzae Rd chromosome. More than 300 clones were sequenced from each end to confirm the overall structure of the genome and identify the six ribosomal operons.*

II.1.2 - State-of-the-art of Shotgun Sanger Sequencing for small genomes decoding.

By the time that we started our own genome sequencing project, in 2006, at least 254 genomes had been fully sequenced and published using the shotgun approach. Eleven years had passed since the 1st genome of *H. influenza has been publicly released (see Fig 9)*. Table 4 illustrates examples of the genomes sequencing during this period of time, where a few illustrating examples are also put in evidence:

Table 4 – Total number of genomes organized by year and domain.
A couple of relevant examples for each are also evidenced.

YEAR	Number of Eubacteria genomes sequenced	Examples of sequenced Eubacteria genomes	Number of Archeal genomes sequenced	Examples of sequenced Archeal genomes	Number of Eukaryotic genomes sequenced	Examples of sequenced Eukaryal genomes
1995	1	*Haemophilus influenzae Rd*				
1996	2	*Mycoplasma pneumonia* *Synechocystis sp. strain*	1	*Methanococcus jannaschii*		
1997	5	*Escherichia coli K-12* *Helicobacter pilori*	3	*Methanogenium frigidum* *Methanobacterium*	1	*Saccharomyces cerevisiae S288C*
1998	5	*Rickettsia prowazekii* *Chlamydia trachomatis*	1	*Pyrococcus horikoshii OT3*	1	*Caenorhabditis elegans*
1999	2	*Deinococcus radiodurans R1* *Thermotoga maritima*	1	*Aeropyrum pernix K1*	1	*Sus scrofer*
2000	18	*Buchnera sp. APS* *vibrio cholerae*	3	*Thermoplasma volcanicum* *Thermoplasma acidofilum*	3	*Arabidopsis thaliana* *Drosophila melanogaster*
2001	20	*Agrobacterium tumefaciens* *C58 Listeria monocytogenes*	4	*Pyrococcus furiosus* *Pyrococcus abyssi*	3	*Homo sapiens* *Encephalitozoon cuniculi*
2002	24	*Buchnera aphidicola* *Mycobacterium tuberculosis*	4	*Methanosarcina mazei* *Methanosarcina acetivarans*	9	*Ciona intestinalis* *Mus musculus*
2003	34	*Yersinia pestis* *Corynebacterium diphtheriae*	2	*Natrialba asiatica* *Nanoarchaeum equitans*	3	*Crytospondium parvum* *Caenohabditis briggsae AF16*
2004	39	*Bdellovibrio bacteriovorus HD100* *Bacilius licheniformis DSM13*	7	*Halobaculum gomorrense* *Haloarcula marismortui*	18	*Homo sapiens* *Rattus norvegicus*
2005	42	*Photobacterium profundum* *Geobacillus jurassicus sp. nov.*	2	*Thermococcus kodakarensis KOD1*	17	*Drosophila pseudoobscura* *Cryptococcus neoformans serotyoe D*
TOTAL	198		26		61	

Our starting point, by that year of 2006 was very encouraging:

- Our team had a ABI 3700 capillary electrophoresis sequencer that could run 96-well plates of sequence reactions for a 2 hour time each plate, with an average of 600 bp per read; We could run between 200 and 300 reactions per day, since the laboratory of STAB VIDA was performing it routinely for its clients;

- We had settled a collaboration with Scylla Ltd, a bioinformatics team in Campinas, Brazil, that had been previously involved in the *Xylella fastidiosa* genome project (Simpson AJ *et al*, 2000);

- A genomic library of *D.gigas* had been already constructed in λ-DASH, using tagged molecular probes designed from known portions of the genome (Thoenes et al, 1994);

- Moreover, the members of the team had been involved in studying several different fragments of *D. gigas* genome, resulting in 14 peer-reviewed publications which are detailed in the Table 5.

Table 5 - Peer-reviewed publications in which the members of the team had been involved, with the intent of studying several different genes of D. gigas.

Year	Publication	Gene	DNA fragment size
1994	Molecular cloning and sequence analysis of the gene of the molybdenum -containing aldehyde oxido-reductase of ***Desulfovibrio gigas***. The deduced amino acid sequence shows similarity to xanthine dehydrogenase. (Thoenes ***et al***, 1994)	mop	mop – 2721 bp
1997	Studies on the Redox Centres of the Terminal Oxidase from ***Desulfovibrio gigas*** and Evidence for Its Interaction with Rubredoxin. (Gomes ***et al***, 1997)	roo rd	Subcloned fragment size - 3600 bp
1999	***Desulfovibrio gigas*** neelaredoxin. A novel superoxide dismutase integrated in a putative oxygen sensory operon of an anaerobe. (Silva ***et al***, 1999)	ORF 1 (mcp) ORF 2 (cheW) nlr	ORF 1 (mcp) – 1949 bp ORF 2 (cheW) – 506 bp nlr – 392 bp
2000	Molecular Cloning of the Gene Encoding Flavoredoxin, a Flavoprotein from ***Desulfovibrio gigas.*** (Agostinho ***et al***, 2000)	flr	flr – 583 bp
2001	Molecular Characterization of Desulfovibrio gigas Neelaredoxin, a Protein Involved in Oxygen Detoxification in Anaerobes. (Silva ***et al***, 2001)	nlr ORF 1 ORF 2	Nlr ≈ 500 bp ORF 1 and ORF 2 not found
2001	Analysis of the ***Desulfovibrio gigas*** Transcriptional Unit Containing Rubredoxin (rd) and Rubredoxin-Oxygen Oxidoreductase (roo) and upstream ORFs . (Silva et al, 2001)	roo rd ORF 1 ORF 2 ORF 3	roo – 1209 bp rd – 158 bp OR F1 – 812 bp ORF 2 – 191 bp ORF 3 – 707 bp
2003	A novel membrane -bound Ech [NiFe] hydrogenase in ***Desulfovibrio gigas.*** (Rodrigues ***et al***, 2003)	echA echB echC echD echE echF	echA – 1943 bp echB – 854 bp echC – 443 bp echD – 377 bp echE – 1076 bp echF – 317 bp
2003	Docking and electron transfer studies between rubredoxin and rubredoxin:oxygen oxidoreductase. (Victor ***et al***, 2003)	rd roo	Not described
2005	Characterization and Expression Analysis of the Cytochrome bd Oxidase Operon from ***Desulfovibrio gigas*** (Mach ado ***et al***, 2005)	Subunit I - cydA Subunit II - cydB	cydA – 1332 bp cydB – 1011 bp
2005	Deletion of flavoredoxin gene in ***Desulfovibrio gigas*** reveals its participation in thiosulfate reduction. (Broco ***et al***, 2005)	flr	4378 bp
2006	***Desulfovibrio gigas*** Flavodiiron protein affords protection against mitrosative Stress in vivo (Rodrigues ***et al***, 2006)	roo	4378 bp

By the time that the laboratory started studying the genes of *D. gigas,* previous evidence pointed out that the *D. gigas* genome size was around 1.6 Mbases, and that it carried two plasmids of around 70 MDal (105 kb) and 40 MDal (60 kb). It also gave a first clue on the very high GC content, around 65% (Postgate *et al*, 1984). We calculated that we would need a coverage of 7x. Fleischmann and collaborators, in 1995, had proposed and used a strategy where they sequenced a number of fragments from both ends to get 6x coverage (Fleischmann *et al*, 1995).

II.1.3 - Description of work - Material and Methods

This work package was done according to the following tasks, distributed between both the ITQB and STAB VIDA laboratories.

II.1.3.1 – Construction of a genomic library of *D. gigas* (shotgun cloning)

This first stage aimed at the preparation of small fragments of genomic DNA from *Desulfovibrio gigas*. The genomic DNA was enzymatically digested to yield fragments of various sizes. Of these, we selected those with sizes between 0.5 kb and 2 kb, for subsequent insertion into a cloning vector, the plasmid pUC19. It was intended, therefore, that the recombinant plasmids represent the entire genome of *D.gigas*. The steps performed where:

- Digestion of genomic *D.gigas* DNA by *Sau*IIIa restriction enzyme, generating fragments of different sizes;

- Restriction analysis of reactions by electrophoresis on agarose gel;

- Selection, extraction and purification of the fragments with sizes between 0.5 and 2.0 kb.

The choice of the size of these fragments took into account both the stability of the cloning vector, such as the extension of the readings performed by the automated sequencer;

- Ligation with the enzyme DNA ligase, of the chosen fragments, into linearized pUC19 cloning vector;

- Obtaining competent *E. coli* cells for transformation;

- Transformation of the competent cells with the recombinant plasmids;

- Selection of positive colonies by the λ-Gal (lacZ) colorimetric method;

- Culture of the positive transformants;

- Extraction and purification of plasmids from a culture of cells positive transformants;

- Analysis of plasmid DNA by agarose gel electrophoresis. Determination of its size, purity and concentration.

II.1.3.2 - Screening of genomic library of cloned *D. gigas* in λ-DASH and subcloning in pUC19:

We have then used the genomic library of *D.gigas* already constructed in λ-DASH, using tagged molecular probes designed from known portions of the genome.

Once identified the plaques that contained complementary sequences to the probes, we proceed to the extraction of its DNA. This was digested and subcloned into minor cloning vectors (plasmid pUC19), which were used to transform competent bacteria. Of the positive clones. The recombinant plasmids of the positive clones were produced. We have done two genomic libraries – one cloning fragments of 2 Kb and a random one by DNA shearing.

The steps performed were:

- Radioactive labelling of probes, designed from portions of the genome of *D. gigas* known and published;

- Extraction and purification of phage DNA of λ-DASH from the phage plaques in which was verified the hybridization to the probes. Analysis of phage DNA by electrophoresis on agarose gel. Determination of size, purity and concentration;

- Digestion of the fragments cloned in phage λ-DASH by the restriction enzymes (eg *Sau*-IIIa), generating fragments of different sizes. Restriction analysis of reactions by agarose gel electrophoresis. Connection with the enzyme DNA ligase of the various fragments to the linearized cloning vector pUC19;

- Obtaining competent *E. coli* cells for transformation;

- Introduction of the recombinant plasmids into the competent bacterial cells;

- Selected positive colonies of transformants by colorimetric method;

- Culture of positive transformants;

- Extraction and purification of plasmids from a culture of positive cells transformants;

- Analysis of plasmid DNA by agarose gel electrophoresis. Determination of size, purity and concentration.

II.1.3.3 – Sequencing of cloned fragments:

The cloned fragments in pUC19 were amplified with the primer pair *universal* and *reverse*. These primers are complementary to the regions flanking the insertion sites of the fragments (MCS - multiple cloning site). The next steps were:

i) purification of the amplification reaction products;

ii) removal of unwanted products;

iii) automated separation of amplified fragments, labelled with fluorescent emitters, capillary electrophoresis;

iv) detection of fluorescent signals;

v) interpretation of the nucleotide sequence generated by automated sequencer. The chemistry used was the Dye-terminator sequencing using Capillary electrophoresis. The dye-terminator sequencing method, along with automated high-throughput DNA sequence analysers, was

used for this part of the sequencing project. The first 15–40 bases of the sequence were of pour quality deteriorating the quality of sequencing traces after 700–900 bases. We used BigDye® Terminator v3.1 Cycle Sequencing Kit.

The thermal cycling and clean up protocols for cycle sequencing have been modified to optimize results using the new formulation of this kit. We ran the samples on 3100 and 3700 Genetic Analyzer, and ABI PRISM 3730 xl Genetic Analyzer.

Figure 10 – *ABI Sequencers used during the Sanger protocol, from left to right: ABI 3700, ABI 3100 and ABI 3730xl (http://www.medwow.com; www.maplewininternational.com).*

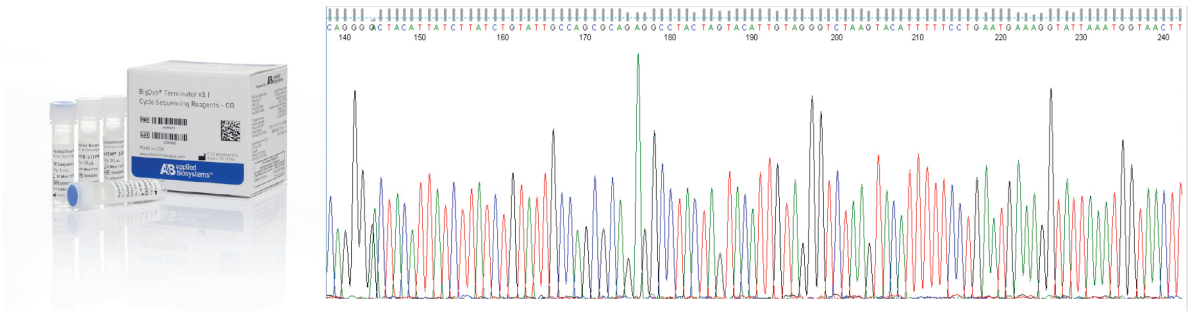


Figure 11 - *Bigdye® terminator 3.1 cycle sequencing kit on the left and a representation of a Sequencing Chromatogram (www.lifetechnologies.com; www.stabvida.com).*

II.1.3.4 - Contig Assembly by the Scylla Bioinformatics team:

The database system Scylla Genome assembled a total of 29.564 reads, with additional 28 sequences of *D.gigas* obtained from Genbank. The assembly procedure was performed using the program phrap (version 990 319). The assembly was carried out in three steps, detailed below:

In the first step, all sequences were processed to obtain a clean first set of contigs. In the second stage, the contigs generated in the previous step were compared using the BLAST program (version 2.2.11). A graph was constructed so that each contig is an apex and each edge indicates the existence of alignment with e-value of 0.0 between contigs represented by the vertices. Then, the process of "reassembling" examined each of the connected components of the graph. In this analysis, all sequences of each connected component in the contigs are provided back to the program phrap. In this implementation the program is configured with less stringent parameters. The third stage was similar to the second one, the graph was however built considering the e-value $1e^{-10}$. The processes of recording the contigs were conducted using a glimmer program (version 3.02) to identify ORFs in each of contig assembly. Software's also used: AMOSCMP software and Bambus software.

II.1.3.5 Contig assembly and annotation by the Fiocruz team:

The assembly of the genome was not progressing satisfactorily and as such we have later switched to the bioinformatics Fiocruz team (BH, Brazil). *Ab initio* assembly was performed using Velvet version 0.7.55 software (Zerbino and Birney, 2008), and the consensus genomic sequence was obtained with Phrap. (http://www.phrap.org/phredphrapconsed.html).

Structural annotation was performed using FgenesB (www.softberry.com), RNAmmer (Lagesen et al. 2007), tRNA-scan-SE (Lowe and Eddy 1997) and Tandem Repeat Finder (tandem.bu.edu/trf/trf.html). Functional annotation was performed by similarity, using public databases and InterProScan analysis (Zdobnov and Apweiler 2001). Protein-coding sequences were manually curated using Artemis (Rutherford et al, 2000).

The results of this team's bioinformatics work on the raw data from Sanger sequencing was gathered with other raw data's analysis (see next sections of this chapter) and published in 2014 (Morais-Silva *et al*, 2014).

II.2 – The novelty of pyrosequencing - getting the first of the three NGS sequencing experiments for GIGASNOMA project

In September 2005, a paper published in Nature by Marcel Margulies and Michael Egholm *et al*, announced the breakthrough of pyrosequencing as the most promising large scale genome sequencing technology (Margulies *et al*, 2005). The paper described the process used to unlock the entire DNA code of the bacteria *Mycoplasma genitalium* in only four hours. In the year of 2008, encouraged with the publications using 454 for sequencing small genomes and hoping to contribute for our project of genome sequencing, we decided to find a sub-contractor for producing a minimum of 25 Mbases of raw data from the DNA of *Desulfovibrio gigas*. At that time, the technology was not mastered at all in any of our laboratories (STAB VIDA and ITQB).

This new technology became an immediate breakthrough for the DNA sequencing and genome sequencing community. Unlike the shot-gun approach, it dismissed the need for a DNA cloning step, and could analyse more than 20 million base pairs at a time. If compared with the shotgun/Sanger approach, where the max output per run are 96 samples, times 700 bp each (around 70 Kb total per run), in the case of pyrosequencing the output was already 300x higher (considering the raw data only). The average base quality claimed was 99%, while in Sanger sequencing it was a bit higher (99.4%) (Margulies *et al*, 2005).

II.2.1 Pyrosequencing: how it began

In 1999, Jonathan Rothberg, the CEO of a small biotech company, named *CuraGen*, found inspiration in the world of microelectronics, where for years chip's processing power and storage capacity had been doubling every year or two. Looking for a gene-sequencing procedure that could be repeatedly miniaturized, he turned to a less powerful technique for dividing genetic material into manageable pieces, which could be analysed simultaneously.

From 1999 to 2005, the team of Dr. Rothberg developed the equivalent to a gene-sequencing "chip" that could analyse more than 20 million base pairs of DNA at a time. They gave it a brand: the 454 sequencer. Its proof of principle was in 2004, when the *M. tuberculosis* study resulted in the identification of the first tuberculosis-specific drug candidate in 40 years and also emphasised the value of the 454 Sequencer for bacterial sequencing applications (Rothberg *et al*, 2008).

It targeted the reduction of cost, complexity and time required for sequencing large amounts of DNA (bacterial & eukaryotic genomes). 454 Life Sciences, based in Branford, Connecticut, started in 2005 to market a tabletop DNA-sequencing machine that worked much faster than mainstream techniques and best machines, made by Applied Biosystems based in Foster City, California, that worked 100 times more slowly. In June 2006, 454 Life Sciences launched a project with the Max Planck Institute for Evolutionary Anthropology to sequence the genome of the Neanderthal, the extinct closest relative of humans. In 2007, the company Roche bought the 454 from Curagen for $155 million in cash and stock, saying at the time the deal would solidify its access to future 454 sequencers and enable it to use the tools for in vitro diagnostic applications. Starting in 2005, prior to the purchase, Roche had been the 454's exclusive distributor. In September 2008 the complete Neanderthal mitochondrial genome was sequenced, establishing the divergence between humans and Neanderthal at 660,000 +/- 140,000 years. The full genome was published later on, in 2010 in Science, using a combination of 454 and Illumina sequencing (Green *et al*, 2006; Green et al 2008; Green *et al*, 2010).

The 454 equipment could be used to sequence any double-stranded DNA and a variety of applications including *de novo* whole genome sequencing, re-sequencing of whole genomes and target DNA regions, metagenomics and RNA analysis. This new method introduced by the 454 pyrosequencing machines proved to be 100 times faster and with a substantial reduction of the cost per base: from 15 US$ cts per base in Sanger Sequencing to approximately US$ 0.015 cts per draft sequence base pair at 10 x coverage (Chan, 2005).

Later on, 454 expanded its output capacity from 20 Mb to much more (≈700 Mb) using the GS FLX Titanium Sequencer. It was introduced in the year 2008, making use of a large-scale parallel pyrosequencing system capable of sequencing roughly 400-600 megabases of DNA per 10-hour run.

II.2.2 - How does pyrosequencing work?

The system relies on fixing nebulized and adapter-ligated DNA fragments to small DNA-capture beads in a water-in-oil emulsion. The DNA fixed to these beads is then amplified by PCR. Each DNA-bound bead is placed into a ~29 μm well on a PicoTiterPlate, a fibre optic chip. A mix of enzymes such as DNA polymerase, ATP sulfurylase, and luciferase are also packed into the well. The PicoTiterPlate is then placed into the GS FLX System for sequencing.

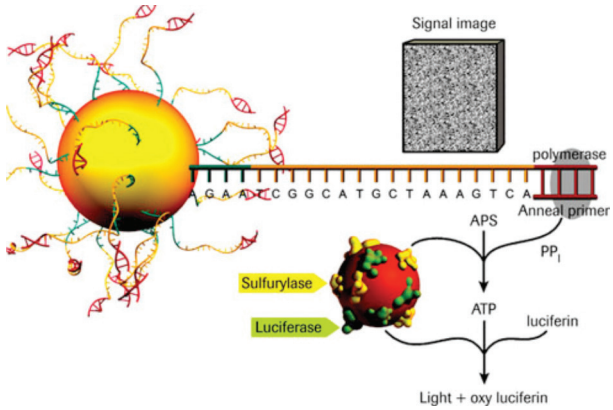


Figure 12 – *Chemical Principle of the Pyrosequencing method.*

In the next pictures, 13 to 18, the innovative 454 method (by the year of 2008, when we first used it) is described and illustrated step by step:

Figure 13 - *Sample Input & Fragmentation - DNA is fragmented into many pieces..*

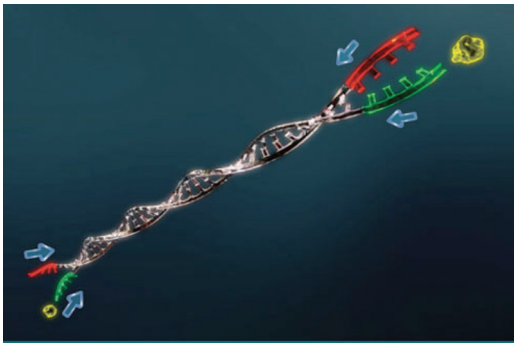


Figure 14 - *ligation of Preparation - Ligate Rapid Library Adaptors to the fragments for use in subsequent purification, quantification, amplification and sequencing steps.*

Random libraries of DNA fragments are generated by shearing the entire genome, and common adapters are added. Afterwards, single DNA fragments are isolated by limiting dilution.

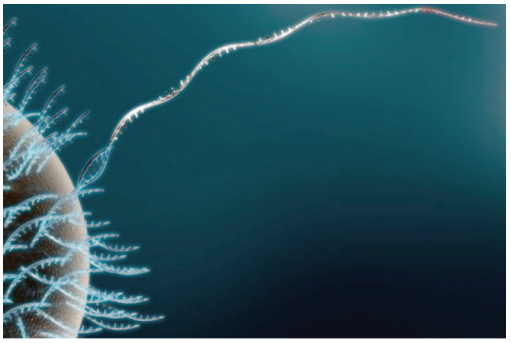


Figure 15 - *One Fragment = One Bead - Attach library to DNA Capture Beads. Each bead carries a unique single-stranded library fragment.*

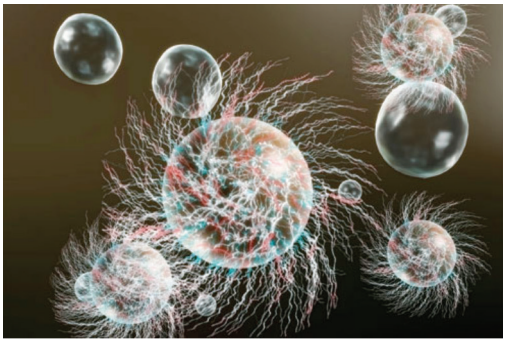


Figure 16 - *The entire emulsion is amplified in parallel, creating millions of clonally copies of each library fragment on each bead.*

These single fragments are captured by their own beads and will be able to perform emulsion PCR: within the droplets of an emulsion, the individual fragment are clonally amplified. This was a great novelty, since it bypassed the need for any kind of subcloning in bacteria, because the templates were handled in bulk within the emulsions.

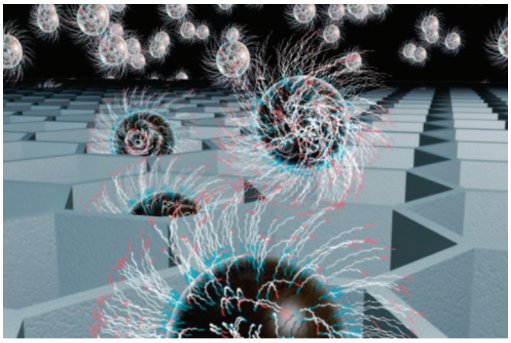


Figure 17 - *One Bead = One Read - the beads are loaded onto the PicoTiterPlate device, where the surface design allows for only one bead per well.*

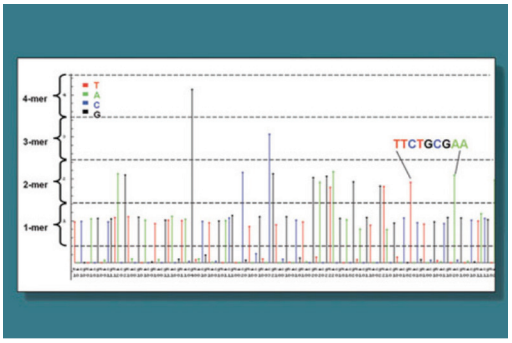


Figure 18 - *454 Sequencing Data Analysis software uses the signal intensity of each incorporation event at each well position to determine the sequence of all reads in parallel.*

Each clonally amplified fragment is then sequenced by synthesis, while a pump drives reagents over the beads to create luminescent signals that show which DNA bases are being used to make copies of DNA. This yields a well-by-well set of sequences.

Tables 6 and 7 make a summary of the evolution of the 454 sequencers, throughout the year 2005 till 2013 (Gilles *et al*, 2011; Droege *et al*, 2008; www.454.com)

Table 6 – Evolution of the Roche sequencers from 2005 to 2009 - commercial and performance characteristics.

Year	2005	2007	2008	2009
Equipment	GS20	GS FLX	Gs FLX Titanium	GS Junior
Average read length	~100bp	~250bp	~350bp	~400bp
Average error rate	Between 0,25% and 0,5%	1,07% (it occasionally rose to more than 50% in certain positions)	99,9% accuracy	Accuracy: Q20 (99%) at 400 bases
Main Bottlenecks	Indel, Mismatches, High cost, low throughput			
Total BIG DATA generated per run	From 20Mbp to 40 Mbp	100 Mbp	700 Mbp	~35 Mbp
Total costs per run (reagents only)	3Mb genome - $12,488 total cost	$8,439 total cost	$6,200	$1500
Costs per Mb of Raw data (reagents only)	4163€ only the sequencing reagents	$84,39 total cost	$12	$22
De novo genome sequencing	Yes			
Application with better performance	Mapping, facilitated *de novo* genome sequencing, short run time			
Examples of *De novo* genome	*Streptococcus suis* (Genbank NC_013698.1); *Clavibacter michiganensis* (Sekizaki *et al*, 2005)	Atlantic Salmon genome (Nicole Quinn *et al*, 2008)	*Jatropha curcas* L. (S Sato *et al*, 2010)	*Pantoea ananatis* (De Maayer *et al*, 2012)

Such an improvement was related to the increase of the length of the individual reads. Most of the new technologies that were entering the market sacrificed length for parallelism and speed even more than 454. Early academic shakedowns suggested that the new Solexa's system, by the year of 2009 produced chains of only 25 base pairs. Rothberg said then, that he expected to increase 454's read lengths to 500 base pairs. By 2012, 454 eventually could generate reads of up to 1000 bp, even surpassing the Sanger technology.

Table 7 – Roche equipment's with the respective commercial and performance characteristics (Gilles *et al*, 2011; www.454.com).

year	2010		2011		2012-2013		
equipment	GS FLX+ Titanium	GS Junior System	GS FLX+ Titanium	GS Junior System	GS FLX+ Titanium ($ 500,000)		GS Junior System
kit					GS FLX Titanium XL+	GS FLX Titanium XLR70	
Average read length	330 bp (up to 500 bp In shotgun libraries)		400-700bp	400bp	Up to 1,000 bp	Up to 600 bp	~400 bp
Average error rate	First 102 positions - 0,534% Full length (500-592 positions) - 1,073%		1%		Consensus accuracy at 15x coverage: 99.99%		99% accuracy
Main bottlenecks	High error rate (most abundant errors – InDels), high cost, low throughput						
Total BIG DATA generated per run	750 MB		500	50	Throughput: 700 MB	Throughput: 450 MB	Throughput : ~35 Mb
Total cost per run	$ 15000		$ 20000	$ 1500	$ 7000		$ 1500
Cost per Mb of raw data	$ 20		$ 12,4	$ 22	$ 10		$ 22
De novo genome sequencing	yes						
Applications with better performance	Mapping, particularly for repetitive regions, facilitated *de novo* genome sequencing, (short run time)						
Example of *de novo* genome	Khoisan and Banatu genomes from southern Africa (Schuster *et al*, 2010)		extremophile crucifer *Thellungiella parvula* (Maheshi Dassanayke *et al* 2011)		*Pantoea ananatis* (De Maayer *et al*, 2012)		

II.2.4 The decision to sub-contract Keygene and Biocant for getting raw data by using 454 pyrosequencing for GIGASNOMA:

In the year of 2008, and very excited with the publications using 454 for sequencing small genomes and since the many thousands of Sanger reads were enough, we decided to find a sub-contractor for producing a minimum of 25 Mb of raw data to add to the *D. gigas* Sanger reads. The companies keygene, in the Netherlands, and Biocant, in Cantanhede, Portugal, were chosen to run the Sequencing experiment, after sending them genomic DNA of *D. gigas*.

II.2.4.1 - Materials and Methods (from sub-contractors)

Genomic DNA was fractionated into smaller fragments (300-800 base pairs) and polished (made blunt at each end). Short adaptors were then ligated onto the ends of the fragments. These adaptors provide priming sequences for both amplification and sequencing of the sample-library fragments. One adaptor (Adaptor B) contained a 5'-biotin tag for immobilization of the DNA library onto streptavidin-coated beads. After nick repair, the non-biotinylated strand is released and used as a single-stranded template DNA library. The DNA library is assessed for its quality and the optimal amount needed for the emulsion PCR is determined by titration.

The DNA library was immobilized onto beads. The beads containing a library fragment carry a single sstDNA molecule. The bead-bound library was then emulsified with the amplification reagents in a water-in-oil mixture. Each bead was captured within its own microreactor where the amplification occurs. This results in bead-immobilized, clonally amplified DNA fragments.

Single-stranded template DNA library beads were added to the DNA Bead Incubation Mix (containing DNA polymerase) and were layered with Enzyme Beads (containing sulfurylase and luciferase) onto a PicoTiterPlate device. The device was centrifuged to deposit the beads into the wells. The layer of Enzyme Beads ensures that the DNA beads remain positioned in the wells during the sequencing reaction. The bead-deposition process intends to maximize the number of wells that contain a single amplified library bead.

The loaded PicoTiterPlate device was placed into the Genome Sequencer FLX Instrument. The fluidics sub-system delivered sequencing reagents (containing buffers and nucleotides) across the wells of the plate. The four DNA nucleotides were added sequentially in a pre-defined order across the PicoTiterPlate device during a sequencing run. During the nucleotide flow, millions of copies of DNA bound to each of the beads are sequenced in parallel. When a nucleotide that is complementary to the template strand is added into a well, the polymerase extends the existing DNA strand by adding nucleotides. Addition of one nucleotide, or more generates a light signal that is recorded by the CCD camera in the instrument. This technique is based on sequencing-by-synthesis and is called pyrosequencing. The signal strength is proportional to the number of nucleotides added; e.g. homopolymer stretches, incorporated in a single nucleotide flow generate a greater signal than single nucleotide. However, the signal strength for homopolymer stretches is linear only up to eight consecutive nucleotides after which the signal falls-off rapidly. The data was stored in standard flowgram format (SFF) files for downstream analysis.

II.2.4.5 - Bioinformatics Method (from sub-contractors)

The produced raw data from the 3 runs performed we analysed, isolated and taken together, at Fiocruz's bioinformatics laboratory (BH, Brazil) as described already in Section II.1.3.5 (page 39).

II.3 – Solexa : Illumina

In March 2008, we decided to collect more sequencing raw data to add to the existent 83.6 Mb so far accumulated. Although it meant a brute coverage of already 23x, in fact the assembly work using different bioinformatics approaches performed by the FioCruz Brazilian team could not close the genome. Too many contigs and too many gaps between them were not yet covered.

Two factors explained this fact: the extremely high GC content present in the genome of the *Desulfovibrio gigas* and the relatively poor quality data obtained from the 454 experiments. In fact, our experience today (April 214) tells us that the minimum raw data necessary is 1 Gigabases for a small bacterial genome, with minimum read lengths of 400 bp. By that time (2008) we thought that 80Mb of raw data for a genome of 3.6Mb was tremendously high. As an alternative to another 454 run, we decided to make a trial on a novel system, known as Solexa, which had been recently acquired (2008) and commercialized by Illumina. Since none of the members of our consortium had previous experience with this technology, we have contacted and subcontracted the HTGU (High Throughput Genomic Unit) of the Washington University, USA. A run on their Genome Analyzer (GA) took a month and cost us 3.549,49€, after we sent them the biological material (10 μg of gDNA in dry ice). Later on, in 2012 we insisted one more time and sent more DNA to the Belgium company Baseclear, Ltd to obtain more sequence raw data from another run on their Illumina's Hiseq 2000.

II.3.1 The Illumina's Genome Analyzer: a brief history of the DNA Sequencing-by-synthesis

In the mid-1990s, while performing experiment that intended to observe the motion of a polymerase at the single molecule level, as it synthesized DNA immobilized to a surface with fluorescently labeled nucleotides, Shankar Balasubramanian and David Klenerman, speculated how the approach that they were using might be used to sequence DNA. In 1997, a new concept was born, the concept of using clonal arrays and massively parallel sequencing of short reads using solid phase sequencing by reversible terminators. This was subsequently referred to as sequencing by synthesis (SBS) as the basis of a new DNA sequencing approach (source http://www.illumina.com, adapted).

Balasubramanian and Klenerman obtained initial seed funding to form Solexa in 1998 (from Abingworth Management), and in 2001, the team's research progress attracted £12 million of funding. Three years later, Solexa acquired Manteia's molecular clustering technology. The amplification of single DNA molecules into clusters enhanced the fidelity and accuracy of base calling, while reducing the cost of the optics on the system through the generation of stronger signal (source http://www.illumina.com, adapted).

In 2005, the method was used for sequencing the 5,300-base-pair virus PHIX174 genome, the same genome that Sanger first sequenced and also an 180,000-base-pair bacterial artificial chromosome. Still in 2005, Solexa acquired the Lynx Therapeutics Company in a reverse merger, and started to work on the transformation of the Solexa prototype into a commercial

sequencing instrument. The first Solexa sequencer, the Genome Analyzer, launched in 2006, had the power to sequence 1 gigabase in a single run. Subsequently, in 2007, Solexa was acquired by Illumina and the technology and instrumentation have now sequenced hundreds of microbial, plant, and animal genomes. At the same time this technology kept evolving, through refinements and optimization, with the proof being the newest generation of Illumina SBS technology-based instruments that can generate over 1 terabase of data per run (source http://www.illumina.com, adapted).

In 2009, by the time all previous plant genome sequences have been derived using traditional Sanger technology, Huang and collaborators took advantage of the long read and clone length of Sanger technology and from the high sequencing depth and low unit cost of Illumina GA technology to sequence the cucumber genome. They were able to generate a total of 26.5 gigabases with high-quality, reaching a 72.2-fold genome coverage, with Sanger reads providing 3.9-fold coverage and the Illumina reads providing 68.3-fold coverage and ranging in length from 42 to 53 bp (Huang *et al*, 2009).

In 21 January 2010, using the Illumina Genome Analyser sequencing technology alone, Ruiqiang and co-workers generated, assembled and drafted the genome sequence for the giant panda, with an assembled N50 contig size reaching 40 kilobases, and an N50 scaffold size of 1.3 megabases. That achievement represented the first fully sequenced genome of the family Ursidae and the second of the order *Carnivora*. They also carried out several analyses that included genome content, evolutionary analyses, and investigation of some of the genetic features underlying the panda's unique biology. The demonstration that the next-generation sequencing technology allowed accurate *de novo* assembly of the giant panda genome, promoted the construction of reference sequences for other animal and plant genomes in an efficient and cost-effective way (Li *et al*, 2010).

According to market share analysis, almost two thirds of all NGS instruments presently in operation have been manufactured by Illumina (Minoche *et al*, 2011).

II.3.2 The evolution of the Illumina technology

The Illumina sequencing technology has been under constant development, relating to instrumentation, signal processing software, and sequencing chemistry, towards the production of more data and longer sequencing reads (Minoche, 2011).

At first, Solexa Genome Analyzer output was 1 Gbase per run, but through improvements in polymerase, buffer, flowcell, and software, in 2009 the output of GA increased to 20 Gbase per run in August (reads of 75 paired ends, also known as "pe"); 30 Gbase per run in October (reads of 100pe); and 50 Gbase per run in December (Truseq V3, reads of 150pe) (Liu *et al*, 2012).

In early 2010, Illumina launched HiSeq series, which used a chemistry that was similar to the Illumina GA series but with a two to five-fold increased rate of data acquisition. Also, instead of using only one camera, the HiSeq innovated by operating with a four camera system that detects the intensities of all four bases simultaneously (Minoche *et al*, 2011).

Initially it had a 200 Gbase per run output, which was improved to 600 Gbase per run and could be finished in 8 days. In 2011 Illumina also launched a bench top sequencer, the MiSeq, which shared most technologies with HiSeq, but was especially convenient for amplicon and bacterial sample sequencing. It could generate 1.5 Gbase per run in about 10 hours with sample and library preparation time included. In 2012, the Hiseq2500 was launched and it was able to reach 1 Tbase per run (Liu *et al*, 2012).

In the ongoing year, 2014, the HiSeq X has been presented by Illumina, which claims a 1.8 Tbase per run output. (source: Ilumina's website)

Table 8 - Evolution of the Illumina sequencers from 2006 to 2014 - commercial and performance characteristics.

Year	2006	2008	2010	2011	2012	2014
Equipment	Genome Analyzer I	Genome Analyzer IIx	HiSeq 2000	MiSeq	HiSeq 2500	HiSeq X
Average read length	75-100 bp	2 x 150 bp	2 x 100 bp	2 × 250 bp	2 × 100 bp	2 × 150 bp
Average error rate		Mostly > Q30	Mostly > Q30	Mostly > Q30	>90% above Q30	
Main Bottlenecks	Cannot resolve short sequence repeats; Substitution errors		Relatively short read lengths	Higher Indel Rates Errors with CG -rich sequences	Relatively short read lengths	Relatively short read lengths
Total RAW DATA generated per run	50 Gb	85 Gb	600 Gb	1.5 Gb	1000 Gb	1600 – 1800 Gb
Costs per Gb of Raw data (reagents only)		$148	$412	$502	$300	$150
De novo genome sequencing	Yes					
Examples of *De novo* genome	*Pseudomonas syringae pv. syringae* B728a (Farrer *et al*, 2009)	*Linum usitatissimum* (Wang *et al* 2012)			BTV -4 strain YTS -4 (Yang *et al*, 2012)	*Aquila chrysaeto (Doyle et al, 2014)*

II.3.3 The experimental method of sequencing-by-synthesis

The Illumina system uses a sequencing-by-synthesis approach in which all four nucleotides are added simultaneously to the flow cell channels, along with DNA polymerase, for incorporation into the oligo-primed cluster fragments. Specifically, the nucleotides carry a base-unique fluorescent label, with the 3'-OH group being chemically blocked so that each incorporation becomes a unique event. An imaging step follows each base incorporation step, during which the flow cell lanes are imaged in three 100-tile segments by the instrument optics at a cluster density per tile of 30,000. After each imaging step, the 3' blocking group is chemically removed to prepare each strand for the next incorporation by DNA polymerase. These series of steps continues for a specific number of cycles, as determined by user-defined instrument settings, which permits discrete read lengths of 25–35 bases. A base-calling algorithm assigns sequences and associated quality values to each read and a quality checking pipeline evaluates the Illumina data from each run, removing poor-quality sequences (figs 19 to 30, from Illumina Website).

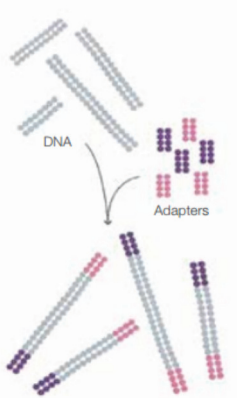


Figure 19 – *Prepare genomic DNA sample: randomly fragment genomic DNA and ligate adapters to both ends of the fragments.*

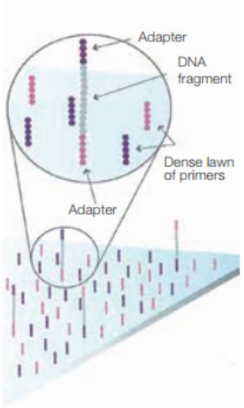


Figure 20 – *Attach DNA to surface: Bind single-stranded fragments randomly to the inside surface of the flow cell channels.*

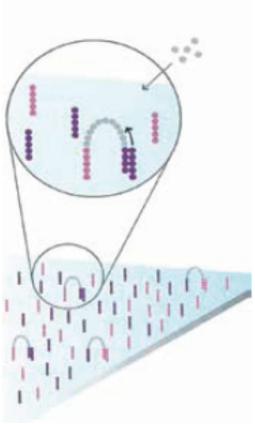


Figure 21 – *Bridge Amplification: Add unlabelled nucleotides and enzyme to initiate solid-phase bridge amplification.*

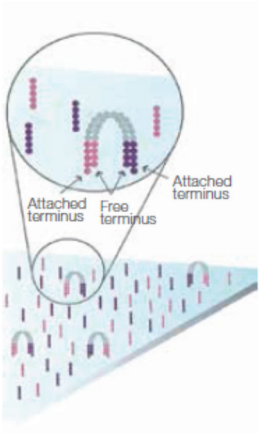


Figure 22 – *Fragments become double stranded: The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.*

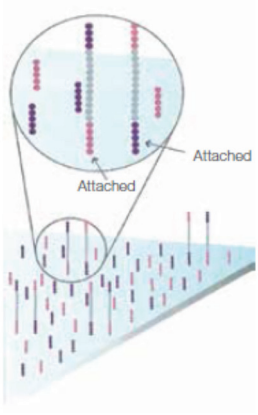


Figure 23 – *Denature the double-stranded molecules: Denaturation leaves single-stranded templates anchored to the substrate.*

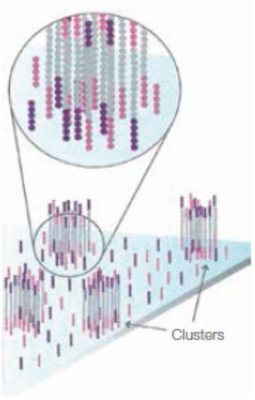


Figure 24 – *Complete amplification: Several million dense clusters of double stranded DNA are generated in each channel of the flow cell.*

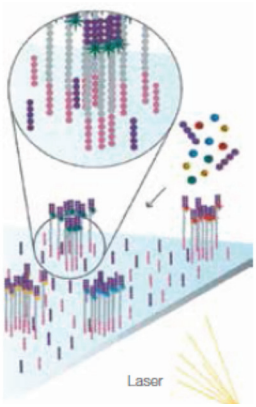


Figure 25 – *Determining first base: the first sequencing cycle begins by adding four labelled reversible terminators, primers and DNA polymerase.*

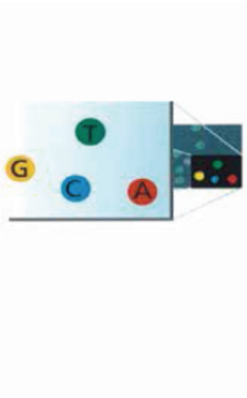


Figure 26 – *Image first base: After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.*

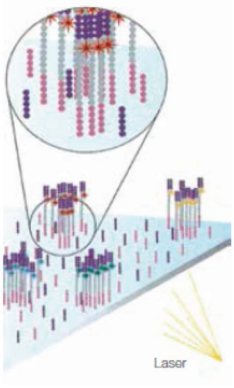


Figure 27 – *Determining second base: the next cycle repeats the incorporation of four labelled reversible terminators, primers and DNA polymerase.*

Figure 28 – *Image second chemistry cycle: After laser excitation, the image is captured as before and the second base is recorded.*

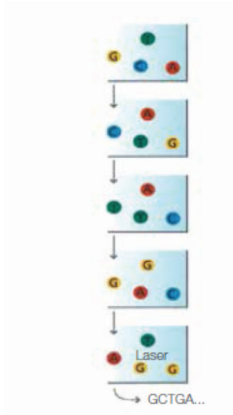


Figure 29 – *Sequencing over multiple chemistry cycles: The sequencing are repeated to determine the sequence of bases in a fragment, one base at a time.*

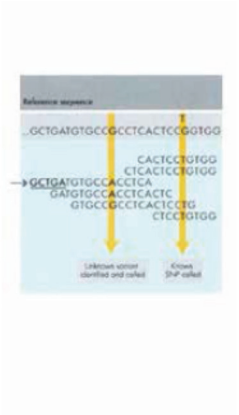


Figure 30 – *Align data: the data are aligned and compared to a reference, and sequencing differences are identified.*

II.3.3.1 - Subcontracting Washington University and Baseclear Ltd

In the year 2008, we sent 10 μg of *D. gigas* DNA to the high-throughput unit of the Washington University, where a run on the Genome Analyzer was performed. Later on, during the year 2012, we sent another 2 μg of the *D.gigas* DNA to our collaborating company in Belgium, Baseclear, where a run on the Hiseq 2000 machine was performed. The obtained raw data were then sent to the bioinformatics team in Belo Horizonte, Brazil.

Figure 31 – *Equipments used for generating raw data with Illumna technology, from left to right: the Genome Analizer and the Hiseq2000 (source: Illumina inc, adapted).*

II.3.3.2 - Raw data processing by the bioinformatics team

The produced raw data from the 2 SBS experiments was analysed, isolated and taken together with the raw data from Sanger Sequencing and the raw data from 454 sequencing, at Fiocruz bioinformatics laboratory, as described in II.1.3.5 (page 39).

II.4 - The Ion Torrent from Life technologies for producing BIG DATA on the *D. gigas* genome.

After gathering raw data from three different approaches – Sanger Sequencing, two 454 runs and two Illumina runs – we were still unable to close and complete the genome sequence of *D. gigas.* Indeed, a total number of 20 gaps were not yet closed.

Meanwhile, STAB VIDA invested on its own NGS platform, buying the PGM machine from Life Technologies, supported on the Ion Torrent methodology, early in the year of 2013.

By February 2013 we tried our first small genome run at STAB VIDA laboratory located in the building of the IMM, Lisbon. The DNA from *D. gigas* had been prepared in Rodrigues- Pousada's laboratory, in ITQB, Oeiras. We needed a successful experiment since it was essesntial to have *D. gigas* genome in a unic contig, and hoped that the Ion Torrent experiment helped us to close it for good.

II.4.1 Analysing the Ion Torrent: a brief history

The Ion Torrent Personal Genome Machine® (PGM™) (Life Technologies Corporation, Grand Island, NY, USA) consists of a sequencing approach based on the measurement of the hydrogen ion release during deoxynucleotide incorporation. Ion PGM was released by Life Technologies at the end of 2010. Until the development of this technology, sequencing was limited by requirements of imaging technology, electromagnetic intermediates and specialized nucleotides or other reagents. To overcome the previously referred limitations a shift based on a non-optical sequencing was pursued (Liu *et al*, 2012; Dark, 2013).

In 21st July of 2011, Rothberg and collaborators reported a DNA sequencing technology, using an integrated circuit that was able to directly perform non-optical DNA sequencing of genomes through scalable and low-cost semiconductor manufacturing techniques (Rothberg *et al*, 2011). Previous attempts to detect both single-nucleotide polymorphisms (SNPs) and DNA synthesis as well as sequence DNA electronically had already been made, yet, *de novo* DNA sequence had not been produced, neither the issue of delivering template DNA to the sensors or the scaling to large arrays had been addressed. After focusing the development of the ion chips, and all the biochemical methods and software needs inherent to *de novo* DNA sequencing, the previously encountered limitations were overcome. The performance of the ion chips and the overall sequencing platform was demonstrated through whole-genome sequencing of three bacterial genomes, Rothberg and collaborators succeeded in the sequencing of all three genomes five-fold to ten-fold in individual runs, using the small ion chip, and being able to cover 96.80% to 99.99% of each genome (Rothberg *et al*, 2011). The *E. coli* genome was sequenced using three consecutively larger ion chips. Here it is a summary of all these 3 genomes:

- *Vibrio fisheri*, with a genome size of 4.2 Mb had a 6.2-fold coverage, with 96.80% coverage, with a total of 26.0 Mb mapped (1.2M well/chip);
- *Rhodopseudomonas palustris*, with a genome size of 5.5 Mb had a 6.9-fold coverage and 99.64% coverage, with a total of 37.8 Mb mapped (1.2M well/chip);

• *Escherichia coli*, with a genome size of 4.7Mb, had an 11.3-fold coverage, with 99.99% coverage, with a total of 47.6 Mb mapped (1.2M well/chip), 169.9 Mb (6.1M well/chip) and 273.9 Mb (11M well/chip).

After the written consent provided by Gordon Moore (the author of "Moore's law") to sequence and publish his genome and resulting findings, the scalability of the chips architecture was demonstrated by using chips with up to 10 times the number of sensors, and producing a low coverage sequence of his genome. After 1,601 runs using the 1.2M well ion chip, 267 runs using the 6.1M well ion chip and 28 runs using the 11.1M well ion chip, Moore's genome sequence had a 10.6-fold coverage, with 99.21% coverage, with a total of 30.2 Gbases mapped. For the first time, 'post-light' genome sequencing of bacterial and human genomes had been performed (Rothberg et al, 2011).

In 20th July of 2011, Mellmann and collaborators published a work, referring to an outbreak of virulent Shiga toxin (Stx)-producing *Escherichia coli* (O104:H4), in Germany, that had caused over 830 cases of hemolytic uremic syndrome (HUS) and 46 deaths since May 2011. In that work they were able to use the Life Technologies Ion Torrent PGM™ sequencer and Optical Mapping - a fully automated single molecule system for creating ordered restriction maps directly from genomic DNA molecules, to perform whole genome sequencing to characterize the outbreak isolate (LB226692) and a historic O104:H4 HUS isolate from 2001. Reference guided draft assemblies of both strains were completed with the newly introduced PGM™ within 62 hours and allowed them to rapidly conclude that, though closely related, the outbreak strain differs from the 2001 strain in plasmid content and fimbrial genes, allowing them to propose a model in which both strains evolved from a common EHEC O104:H4 progenitor (Zhou et al, 2007; Mellmann et al, 2011).

This outbreak practical case allowed to conclude that rapid next-generation technologies facilitate prospective whole genome in a rapid and efficient manner when needed, and as such, may become imperative in diagnosis. Life Technologies was so euphoric with this study of the outbreak that they claimed that it was a matter of time before their Ion Torrent technology would surpass Illumina's the first mover's advantage.

II.4.2 The Ion Torrent Technology

The Ion Torrent, being classified as the first PostLight™ sequencing technology is innovative due to the its basis that relies on the creation of a direct interaction between the chemical and the digital information, allowing fast, simple and scalable sequencing (Shekhar et al, 2011). The Ion PGM uses a semiconductor technology: the principle is that when a nucleotide is incorporated into the DNA molecules by the polymerase a proton is released as a byproduct and, as such, through detections of changes in the pH, the PGM recognizes if the nucleotide was added or not (Shekhar *et al*, 2011; Liu *et al*, 2012).

The ideal in simplicity would be to detect synthesis directly, via a transistor-based sensor, without the use of labelled nucleotides. The basic concept is related to sequencing-by-synthesis with electrochemical detection of synthesis, through its own sensor, which in turn are organized into a parallel sensor array on a CMOS chip (Merriman *et al*, 2012).

DNA fragments with specific adapter sequences are linked and clonally amplified by emulsion PCR on the surface of 3μm diameter beads, known as Ion Sphere Particles, and then the tem-

plate beads are loaded into proton-sensing wells, which are fabricated on a silicon wafer, and sequencing is primed from a specific location in the adapter sequence. Each of the four bases is introduced sequentially, and therefore, when that specific base is incorporated there will be proton release. Beneath the wells there is an ion sensitive layer and beneath that a proprietary ion sensor, that allow the consequent detection of the alteration of the pH, with the signal being proportional to the number of bases added, without scanning, cameras or light (Sekhar *et al*, 2011; Quail *et al*, 2012).

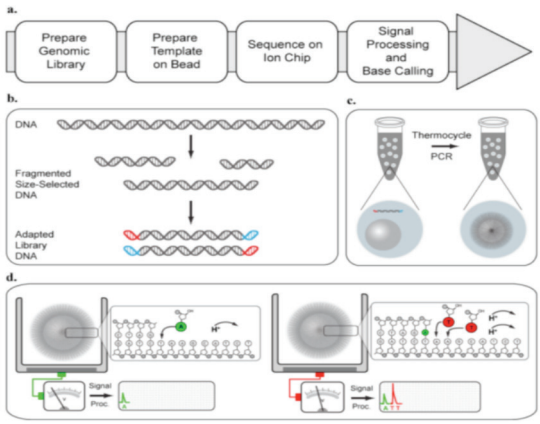


Figure 32 - *http://www.genomics.cn*

The fundamental chemical byproducts of synthesis, in this case generated after polymerase incorporation, are the release of pyrophosphate cleaved from the incorporated nucleotide, the well-known basis for pyrosequencing, and also the release of a hydrogen ion (H^+) from the 3' OH incorporation site on the growing strand. Thus, a sensor capable of detecting H^+ would be adequate as a mean of direct detection of nucleotide incorporation. The transistor-based detection of the H^+ is the functional component of widely used solid-state pH meters and as such it is a well-established technology. This classical transistor device is known as a pH-sensitive field effect transistor (pHFET) (Merriman *et al*, 2012).

The integrated circuit consists of a large array of sensor elements, each with a single floating gate connected to an underlying ISFET. To ensure sequence confinement 3.5mm diameter wells are used, formed by adding a 3-mm-thick dielectric layer over the electronics and etching to the sensor plate. The high-speed addressing and readout are accomplished by the semi-conductor electronics integrated with the sensor array, beneath each microwell the Ion-Sensitive Field Effect Transistor (ISFET) detects the pH change as a result of each proton release and a potential change (DV) is recorded as direct measurement of nucleotide incorporation events. Due to the change in voltage with the number of nucleotides incorporated at each step being scalable, Ion Torrent's sequencing technology has an inherent capacity of repeating calls (Niedringhaus *et al*, 2011; Rothberg *et al*, 2011; Hui, 2012). The general approach could certainly accommodate a removable terminator chemistry to force no more than a single incorporation each time, but the simplest chemistry is to use natural, unmodified nucleotides, which favours better enzyme activity, hence long reads, as well as lower cost reagents (Merriman *et al*, 2012).

The raw voltages need to be changed into base calls, to do so, a signal-processing software converts the raw data into measurements of incorporation in each well for each successive nucleotide flow using a physical model that takes into consideration diffusion rates, buffering

effects and polymerase rates applied to incorporation signals are extracted.

The signals are corrected for phase and signal loss and corrected base calls are generated for each flow in each well to produce the sequencing. Next, each read is sequentially passed through two signal-based filters to exclude low-accuracy reads.

• The first filter measures the fraction of flows in which an incorporation event was measured. When this value is unusually large (greater than 60% of the first 60 flows) the read is not clonal.

• The second filter measures the extent to which the observed signal values match those predicted by the phasing model. When there is poor agreement (median absolute difference more than 0.06 over the first 60 flows) between the two, it corresponds to higher error rates.

Finally the quality per-base values are predicted using an adaptation of the Phred method quantifying the concordance between the phasing model predictions and the observed signal (Rothberg *et al*, 2011).

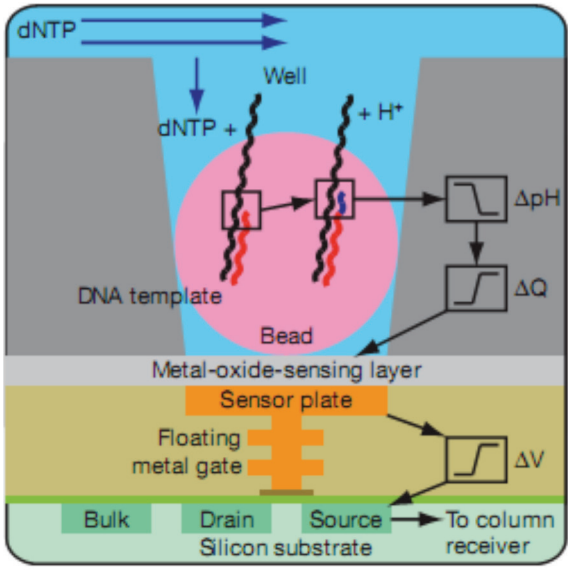


Figure 33 - *Representation of a well and bead containing DNA template, and the underlying sensor and electronics.*

II.4.3 The evolution of the Ion Torrent

The fundamental sequencing capacity of the chips is set by the number of sensors present in the array, which restricts the maximum number of reads producible per run. To increase throughput, the Ion Torrent sequencing chip makes use of a highly dense microwell array in which each well acts as an individual DNA polymerization reaction chamber containing a DNA polymerase and a sequencing fragment. Intrinsically, since the fundamental array architecture is scalable, this can be increased radically, and this has been undergoing rapid progression on successive chips: the 314, 316, 318 chips provided 1.2, 6.3, and 11.3 million (M) fluid addressable wells on the, respectively, and the Proton I and Proton II chips, have 165 and 660 M fabricated wells respectively (Niedringhaus *et al*, 2011; Merriman et al, 2012). Ultimately, Ion Torrent seeks to "democratize" sequencing by offering the first reasonably priced ($50K) benchtop-scale, high-throughput sequencing machine (Niedringhaus *et al*, 2011).

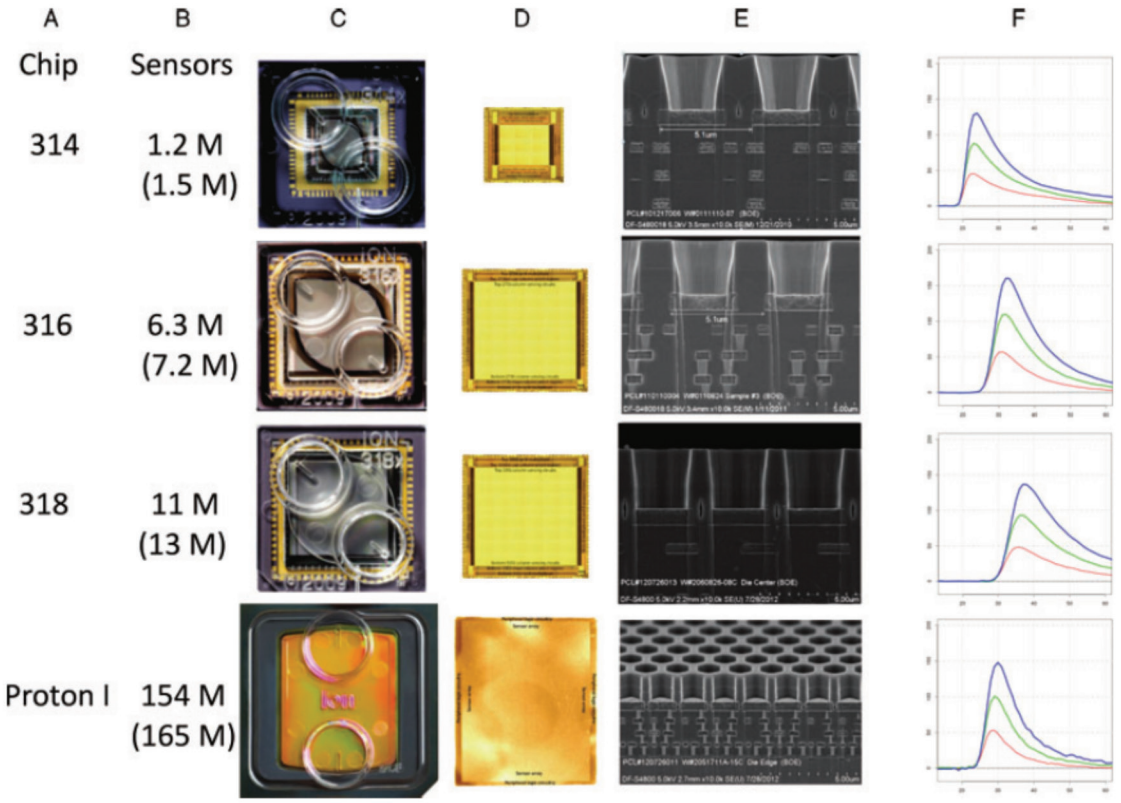


Figure 34 – *Chip scaling. This figure shows the Moore's Law style scaling of successive chip generations. Column (A) is the chip name, column (B) is the number of fluidic addressable wells/sensors on the chip (in millions), with the total number of wells/sensors fabricated on the chip in parentheses, column (C) is an image of the packaged chip, column (D) shows the relative size of the unpackaged, cut CMOS die, and of the sensor array area within the chip, and column (E) shows electron micrographs of sections through the sensor array (individual microwells and underlying electronics visible), with all images shown to scale across the chip series. (Merriman et al, 2012).*

II.4.4 The experimental work performed

After receiving 2 μg of genomic DNA of *D. gigas* extracted and purified at ITQB, STABVIDA's laboratory did the wetlab experimental protocol. The work flow consisted of four major steps: library construction, template preparation, sequencing and analysis.

II.4.4.1 - Wetlab phase of the work

The first step in the workflow was to generate a library of DNA fragments flanked by the Ion Torrent adapters. This was done by ligating the adapters to the PCR products. As fairly standard the DNA was fragmented it to a uniform size (generally 200-400b) and then the sequence adapters were added.

The library fragments were then clonally amplified onto the Ion Sphere™ particles. The fragments generated during the library preparation were attached to beads and amplified using emulsion PCR (emPCR). Beads coated with complementary primers were mixed with a dilute aqueous solution containing the fragments to be sequenced along with the necessary PCR reagents. At that point the solution was mixed with oil to form an emulsion of microdroplets. The concentration of beads and fragments needed to be kept low enough such that each microdroplet contains only one of each. Clonal amplification of each fragment was then performed within the microdroplets. Following amplification the emulsion was 'broken' (by organic extraction and centrifugation) and the amplified beads were enriched in a glycerol gradient (with the unamplified beads becoming pelleted).

The Ion Sphere™ particles coated with template were then applied and deposited in the Ion chip wells by a short centrifugation step and afterwards the chip was placed on the PGM. As previously described what really differentiates Ion Torrent's systems is the sequencing technology. It is based on the standard pyrosequencing chemistry, a form of 'sequencing by synthesis' whereby individual bases are introduced one at a time and incorporated by DNA polymerase. However, unlike other platforms based on pyrosequencing, rather than measuring light released from chemiluminescent reagents, the Ion Torrent system measures the direct release of H+ (protons) from the reaction.

II.4.4.2 - Bioinformatics on the produced raw data

As the Ion Torrent systems generate standard output files like FASTQ, data analysis is generally straightforward. The analysis was performed by at Fiocruz Institute (Belo Horizonte, Brazil)who added this data to the previous gathered data and has performed the assembly as described in II.1.3.5 .

II.5 Results and discussion

This next section summarizes the overall results of the raw data of DNA sequences obtained from the 4 methods described before: Sanger, *454 Pyrosequencing*, Sequencing-by-synthesis (Illumina) and Ion Torrent (Life Technologies).

The discussion refers mainly to the technological approaches, and not on the assembly and annotation of such raw data and genomic information of *D. gigas*, which is further presented in chapter IV.

II.5.1 - Comparison of the obtained raw data

During 8 years, from 2006 till 2014, a total of 1.5 Gb of DNA base pairs were gathered, from a bit more than 23 million reads of DNA sequences. As the project was advancing, new technological approaches were being adopted by DNA laboratories, and we must say that we have tried all of them.

From 23 thousand capillary electrophoresis reads, we reached 20 million reads from Illumina's platforms Genome Analyser and Hiseq 2000. All this data was used for assembling the genome bioinformatics team at Fiocruz laboratory, at Belo Horizonte, Brazil. From each new experiments, the raw data was being sent to this team, but the number of contigs was still high, until the very last experiment where another 360Mb of DNA data closed all gaps and gave us the two expected contigs: one for the bacterial chromosome, and the other for the bacterial plasmid. These results are presented and discussed in Chapter IV. Table 9 summarizes the raw data numbers:

Table 9 : A summary of all raw data that were obtained during 8 years, necessary for closing the genome of *Desulfovibrio gigas*. A total of 1.5 Gb was obtained and an average coverage of 415.7 for each base of the DNA of D. gigas was necessary in order to assemble and close the bacteria's genome.

	Total number of reads	Average size of reads (bp)	Total number of bases (Mb)	Average coverage for the expected 3,6 Mb of the genome	Where the sequencing run was performed
Shotgun Sanger Sequencing (capillary electrophoresis on ABI machines)	22,857	1038	23.7	6.5	STAB VIDA (Pt)
Primer Walking Sanger Sequencing (capillary electrophoresis)	1,102	1025	1.2	6.9 (considering cumulative to the above Sanger reads)	STAB VIDA (Pt)
454 pyrosequencing (Roche) – 1st Trial	459,801	96	46	12.8	Keygene (Nl)
454 pyrosequencing (Roche) – 2nd trial	275,549	96	26.5	7.4	Biocant (Pt)
SBS – Solexa trial – Genome Analyzer	2,390,015	32	76.5	21.2	High-throughput unit of Washington University (USA)
SBS – Illmunina Hiseq 2000	17,777,675	54	959.9	266	Baseclear (NL)
Ion Torrent – PGM Life Technologies	2,085,311	174	362.8	100	STAB VIDA (Pt)
Total for completing the genome of *D. gigas*	23,012,310	(from 30 to 1038 bp)	1,496.6	415.7	

II.5.2 – Discussion of the Pyrosequencing: the End of 454

Our experience with 454 technology proved that it was not yet ready to replace Sanger sequencing method for *de novo* assembly of genomes, when we performed the experiments. The key issues were still short read lengths by that time of year 2009 (100 bp on average), the lack of paired end reads, and the lack of accuracy of individual reads, particularly in regions where homopolymers were observed. Besides, short read lengths made it impossible to span repetitive genomic elements. In addition, the lack of paired end reads information for each DNA fragment limited assembly to contigs separated by coverage gaps or repetitive elements, such that larger scaffolds are required for high-quality draft sequence and gap closure was difficult to achieve. Given these limitations, we have found that, for *de novo* genome sequencing, the 454 platform was better when used as a complement to, rather than a replacement of, the Sanger sequencing methods.

We were also able to realize that massive-sequencing methods cut DNA into much shorter snippets (in this case, 250 bases) for decoding than the old method used by us for the Sanger library construction. Shorter snippets make not only reassembly technically more challenging but also result in harder probing of areas of the genome that have large repeating sequences.

Used together, the two technologies could improve the efficiency of whole-genome sequencing. However the 454 technology on its own was not seen as a replacement for the older technology. By the year of 2010, the combination of both available raw data (Sanger and 454) was still not enough to close and solve the full *D. gigas* genome. In 2014, we used the two 454 raw data first isolated and then taken together, in order to assess its usefulness.

When first introduced, 454 equipment was a highly disruptive innovation: a vastly more expensive instrument running on highly cost consumables (we paid 40.000 € for the 3 runs of 454, by that time of 2007, substantially increasing the cost of this genome) and it generated short (at most 100 bp) reads of inferior quality. The advantage was the massive number of such reads per run in a relatively short time.

The problems started when Solexa showed up and positioned itself as an alternative to 454, especially when it was bought by Illumina. Although at the beginning the average read length of Solexa was only 25 bp, every time a new chemistry was developed and released to the users, the read length increased. Roche 454 was losing competitiveness in specific applications, until all of them were tagged by the competitor.

When I started to write of this thesis on the 4th trimester of 2013, I received an e-newsletter where Roche announced that they were planning to shut down the 454 sequencing business in mid-2016.

The quitting of 454 and NGS market came only after several attempts to stay in the race:

- in 2010 Roche forged alliances with IBM and DNA Electronics, but these alliance did not resist long time;

- in early 2012, Roche made an hostile bid for Illumina Inc that eventually reached a price tag of $6.7 billion (around 5000 million of euros) – a bid that was rejected by Illumina, for the surprise of many (including myself).

In fact, Roche could always claim its 2 key advantages: fast sequencing and comparatively long read length (from the initial average of 100 bp, 454 could lately produce average reads of 1000 bp – a 10x improvement!).The two major disadvantages were not solved out in time:

expensive machines and high cost per base. Thus, the advantages shaded out when Illumina released its 300 bp chemistry and even Life Sciences, a new kid on the block (actually, a very old kid that arrived very late) was getting average 400 bp read length on its Ion Torrent platform.

II.5.3 – Discussion of the Illumina Sequencing

Existing studies reporting Illumina's data evaluation discovered several biases, that is, non-random distribution of the reads in the sequenced sample over the reference, reported for the Genome Analyzer I (GA), and a non-random distribution of errors reported for the Genome Analyzer II (GAIIx) (Minoche *et al*, 2011).

The sequencing strategy generates short reads with the most common error type being substitutions. The base-call error rate increases with read length owing to 'dephasing noise'. Furthermore, over-representation of GC-rich regions and under-representation of AT-rich regions have been described (Su *et al*, 2011).

Preferences of certain substitution errors and sequence context are also described as for example, wrong base calls frequently preceded by base G, frequencies of base substitutions varying by 10- to 11-fold, with A to C conversions being the most frequent error. These errors can give rise to false positive or wrong consensus sequences as such might have profound implications on the interpretation the results. Indeed a non-random read distribution can bias profiling of transcripts hampering the detection of sequence polymorphisms in regions of low sequence coverage (Minoche *et al*, 2011).

The Sequence-specific Errors (SSE) are a potential cause of false single-nucleotide polymorphism (SNP) calls and significantly obstructs de novo assembly. Various researchers agree that the quality of the Illumina sequencer reads are significantly lower in the later cycles, with quality decreasing noticeably after 50 cycles. It is possible that it will become more prone to SSE that can be defined as the positions where sequence-specific interference of base elongation in the cyclic reversible termination induces a drastic lowering of base call quality. It will therefore increase the probability of miscalls triggered by two major sequence patterns (Nakamura *et al*, 2011, Liu *et al*, 2012):

- inverted repeats
- GGC sequences

Sequencing on the HiSeq or MiSeq instruments requires heterogeneous base composition across the population of imaged clusters. In order to sequence monotemplates, it becomes necessary to significantly dilute or mix the sample with a complex genomic library to enable registration of clusters. And it's also referred that many times pooling of large numbers of samples is required to achieve lowest costs (Liu *et al*, 2012; Quail *et al*, 2012).

Illumina sequencing has matured to the point where many applications have been developed on the platform exhibiting the astonishing sequence throughput as a major advantage this is , contributing to a substantial informatics challenges that allowed the development of an entire generation of novel algorithms such as Maq, BWA, Novoalign, Bowtie, among others, aiming at addressing the analysis challenges of Illumina sequencing. The Illumina is featuring a very large output and very low reagent cost (Liu *et al*, 2012; Koboldt *et al*, 2013).

II.5.4 - About the Ion Torrent

One of the main problems associated with the Ion Torrent technology is the biased coverage on extremely AT-rich sequences, opposing to what happens in CG-rich residues which yield an accurate coverage. Also, techniques involving manipulation on flow cell, such as FRT-seq and OS-Seq will be impossible using semiconductor techniques. The low average fragment length (100 – 200 bp) also makes *de novo* assembling harder and probably less accurate (Quail *et al*, 2012).

When referring to sequence quality, Yergeau and collaborators, in a study on 2012, consider the Ion Torrent very stable, accurate and useful to microbiology studies, since they obtained almost interchangeable results between the 454 and the Ion Torrent, The Ion torrent is standing out due to its low cost and rapid output. The Illumina HiSeq 2500 (high throughput) stands out to be more accurate, with over 80% above Q30, and yielding around 600Gb per run. Nevertheless in order to generate such data the run time is over 250 hours and the mean read length of 2x100 bp. When considering the Illumina MiSeq although the reported accuracy is similar, only 27 hours are needed, to generate around 8 Gb with a mean read length of 2x250. Opposing to the Illumina equipment, the Ion Torrent despite being able to generate less information, from 30 Mb to 2 Gb, the time needed to yield such result is between 2.3 to 7.3 hours, with a reported accuracy of over 90% above Q20 and mean read length of 200-400 bp. Both technologies require a very low amount of starting material, around 100 μg (Dark, 2013).

II.5.5 – As a final summary

Table 9 summarizes the today's landscape of NGS technologies, and gives a glance on the near-future ones. Depending on the application, there is a more suitable technology, or a combination of two (or more), that should be used in order to get the best results.

The quality of the produced raw data, the error rate and the price have differences, and what is valid today will most likely not be true in the years to come.

Massive sequencing is here to stay. When this genome adventure started, sequencing a gene could take one full year of hard work.. At present, DNA can facilities and private genomic companies, like STABVIDA, are able to perform, in average, one genome each week.

Table 10 - Overview of second generation sequencing machines
(TBA means "to be announced")

	Roche	Illumina				Life technologies		PacBio	Oxford Nanopore	Qiagen
	454 GS FLX+	MiSeq	NextSeq 500	HiSeq 2500	HiSeq X Ten	PGM	Proton	RS II	GridION	GeneReader
Price (US $)	500k	125k	US	740k	1m	50k	149k	700k	?	~100k
Output per run	700 Mb	15 Gb	120 Gb	1 Tb	1.8 Tb	2 Gb	10 Gb	375Mb	?	12 Gb
Time per run (Paired-end run):	23 h	65 h	29 h	6 d	< 3 d	7 h	4 h	3 h	?	TBA
Output per day	700 Mb	~5 Gb	~100 Gb	167 Gb	600 Gb	~2 Gb	~20 G b	~1 Gb	?	TBA
Read length (bp)	1,000	2x300	2x150	2x125	2x150	400	200	8,500	10,000	100

As of today, the biggest competition in NGS technologies seems to be between Illumina Inc and Life Sciences Inc (recently bought by Thermo Inc), but it seems that much more is yet to come.

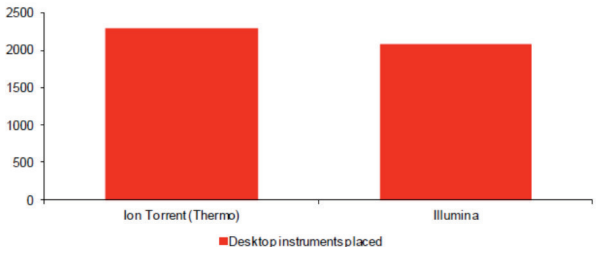


Figure 35 – *Number of desktop sequencing instruments placed. The Ion Torrent has slightly larger market share of desktop instruments placed. However, it's estimated that desktop sequencing represents less than 30% of the total next generation sequencing spend today as high-throughput instruments like Illumina's Hi-Seq remain the sequencing workhorses.*

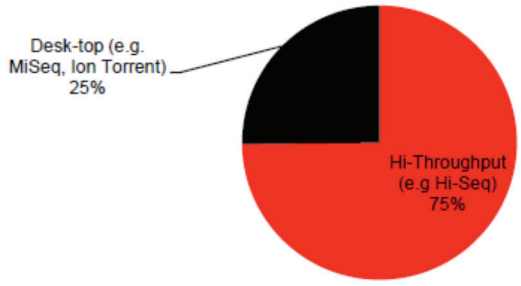


Figure 36 – *High-throughput revenues of consumables spend in 2013.*

Chapter III - In-house Bioinformatics analysis of all raw datas obtained by 4 different sequencing methods:

Sanger, 454, Illumina and Ion Torrent

Preface

Biology faces a new renaissance era, by adopting the BIG DATA on its research, and integrating it into scientific knowledge. For such, bioinformatics is the new science that bridges both the BIG DATA and the knowledge from biological research.

Many were the questions arising from the project of sequencing the full genome of *Desulfovibrio gigas*: why we thought initially that a 7-10x coverage was enough and ended up with the need of almost 500x coverage? Which technological platform performed better, so that we can choose better in future projects? How do we link our knowledge of the biology of *D. gigas*, with the needed knowledge on its DNA sequence?

By adding the 4 types of raw data (Sanger, 454, Illumina and Ion Torrent), was used as reference genome in the exercise of mapping the data against reference, the assembled genome of *D. gigas,* deposited in NCBI database by Rodrigues-Pousada in 2013 (ref CP006585 and CP006586).

Acknowledgments: Besides the author, important contributions for this chapter came from STABVIDA's bioinformatics team, namely Paulo Oliveira and Magda Lewinska.

"*The roots of education are bitter, but the fruit is sweet*"

Aristotle

III.1 - Choice of parameters for replicating the *De novo* assembly of *D.gigas* genome

CLC genomics software was used to assemble the raw data obtained from all sequencing platforms. CLC Genomics Workbench, developed by CLC Bio can be described as a comprehensive and user-friendly analysis package for analysing, comparing, and visualizing sequencing data. This software allowed a contig assembly, as efficiently as possible. After some tries, for this *de novo* assembly, the parameters set for running the software were:

Table 11 - Parameters set for the ***de novo*** assembly on CLC for each of the different origins of raw data.

		Sanger Raw Data	454 Raw Data	Illumina Raw Data	ion torrent Raw Data
Mapping mode	"Map reads to contigs" tool creates a *de novo* object that can be edited	Map reads back to contigs (slow)	Map reads back to contigs (slow)	Map reads back to contigs (slow)	Map reads back to contigs (slow)
Update Contigs	The contigs generated by the de novo assembly are used as references that the reads used for the assembly input are mapped back to	Yes	No	Yes	Yes
Minimum contig length	Contigs below this length will not be reported.	900	200	1000	500
Perform scaffolding	Estimation of gap regions between contigs, with contigs in the same scaffold being outputted as one large contig with Ns between them	Yes	Yes	Yes	Yes
Auto-detect paired distances	Automatic calculus of the distance between the pairs.	Yes	Yes	Yes	Yes
Mismatch cost	The cost of a mismatch between the read and the reference sequence.	2	2	2	2
Insertion cost	The cost of an insertion in the read (causing a gap in the reference sequence)	3	3	3	3
Deletion cost	The cost of having a gap in the read.	3	3	3	3
Length fraction	Minimum length fraction of a read that must match the reference sequence.	0.8	0.8	0.8	0.8
Similarity fraction	Minimum fraction of identity between the read and the reference sequence.	0.8	0.8	0.8	0.8
Create list of un -mapped reads	Creates a list of un -mapped reads	Yes	Yes	Yes	Yes
Word size (bp)	Length of the sub -sequences used to create the Bruijn graphs	19	16	22	19
Bubble Size	Length of the bifurcation in the Bruijn graphs where a path furcates into two nodes and then merge back into one	611	89	50	415
Automatic bubble size		Yes	No	No	No

III.2 - Choice of parameters for mapping all the reads of each raw data, against the reference genome of *D. gigas*.

The raw data obtained from all sequencing platforms and shotgun method was used for a "re-sequencing" exercise, in order to evaluate how complete (or incomplete) was each of the mentioned raw data. Although this raw datas were obtained in the years 2005 til 2013, and the *D. gigas* complete genome was deposited in the NCBI database on the year 2013, this task performed in early 2014, gave us an idea, *a posteriori*, of the performance of each of the sequencing methodologies.

Reference: *Desulfovibrio gigas* DSM 1382 =ATCC 19364		Year 2013
NCBI reference:	CP006585 - chromosome (length 3,693,999 bp)	
	CP006586 – plasmid (length 102,023 bp)	

The obtained contigs from each of the different raw data were mapped against the referred deposited genome of *D. gigas* using the CLC Genomics Workbench, developed by CLC Bio. The following parameters were identified as the most efficient, and consequently the analysis was performed using them:

Table 12 - Parameters set for the mapping to reference on CLC for each of the different origins of raw data.

		Sanger Raw Data	454 Raw Data	Illumina Raw Data	Ion Torrent Raw Data
Masking mode	Mechanism where parts of the reference sequence are not considered in the mapping.	No masking	No masking	No masking	No masking
Mismatch cost	The cost of a mismatch between the read and the reference sequence.	2	2	2	2
Insertion cost	The cost of an insertion in the read (causing a gap in the reference sequence).	3	3	3	3
Deletion cost	The cost of having a gap in the read.	3	3	3	3
Length fraction	Minimum length fraction of a read that must match the reference sequence.	0.8	0.8	0.8	0.8
Similarity fraction	Minimum fraction of identity between the read and the reference sequence.	0.8	0.8	0.8	0.8
Global alignment	Mapping is done with local alignment of the reads to the reference.	No	No	No	No
Non-specific match handling	Non-specific match reads will not be included in the final mapping.	Ignore	Ignore	Ignore	Ignore
Output mode	The read mapping can either be stored as a track or as a stand-alone read mapping	Create reads track	Create reads track	Create reads track	Create reads track
Create report	Generates a summary report	Yes	Yes	Yes	Yes
Collect un-mapped reads	Creates a list of un-mapped reads.	Yes	No	No	No

III.3 Raw data obtained from all sanger reads, before and after trimming

In the next section, what is presented is a description of the raw data consisting of 23,959 sequencing reads obtained from the equipment's ABI 3700, 3100 and 3730xl. The raw data is presented for the two libraries and for primer walking reads, as summarized in Table 13. The most striking result is the reduction on the reads average size, due to trimming process (getting rid of non-sense sequence) from 1000 bp on average to 600 bp.

Table 13 – Description of the raw data before and after trimming for the libraries obtained with Sanger Sequencing.

Description	First Library	Second Library	Primer Walking
Clones obtained from the 1st libraries	15,940	6,917	1,102
Average read length (bp)	1,038	1,025	1,007
Average trimmed read length (bp)	621	597	385
Number of not trimmed reads	0	0	0
Number of reads after trimming	15,522	6,056	958
Successful (%)	97.38%	87.55%	86.93%

Overall, 22,536 high quality reads (i.e. after trimming), with an average size of 603 bp, constitute the trimmed raw data obtained from Sanger Sequencing method. Taking into account the size of 3.6 Mb, this would represent an average coverage of x3.8 times the genome, as illustrated by Table 14.

Table 14 – Number and Size of reads after trimming and expected coverage and genome size.

Total number of high quality reads	22,536
Average size of reads (bp)	603
Total number of bases	13,598,222
Expected genome size (Mb)	3.6
Average coverage	3.8

Figure 37 shows the results of the trimming data process. An ambiguity trimming that only allowed the presence of 4 ambiguous nucleotides was performed. As it is visible, there occurs a significant reduction of the read lengths, with the longest reads being reduced to half of its initial length. Within a total of 23,959 inputted reads, 20,658 were trimmed; 3,301 weren't trimmed and there were no discarded reads.

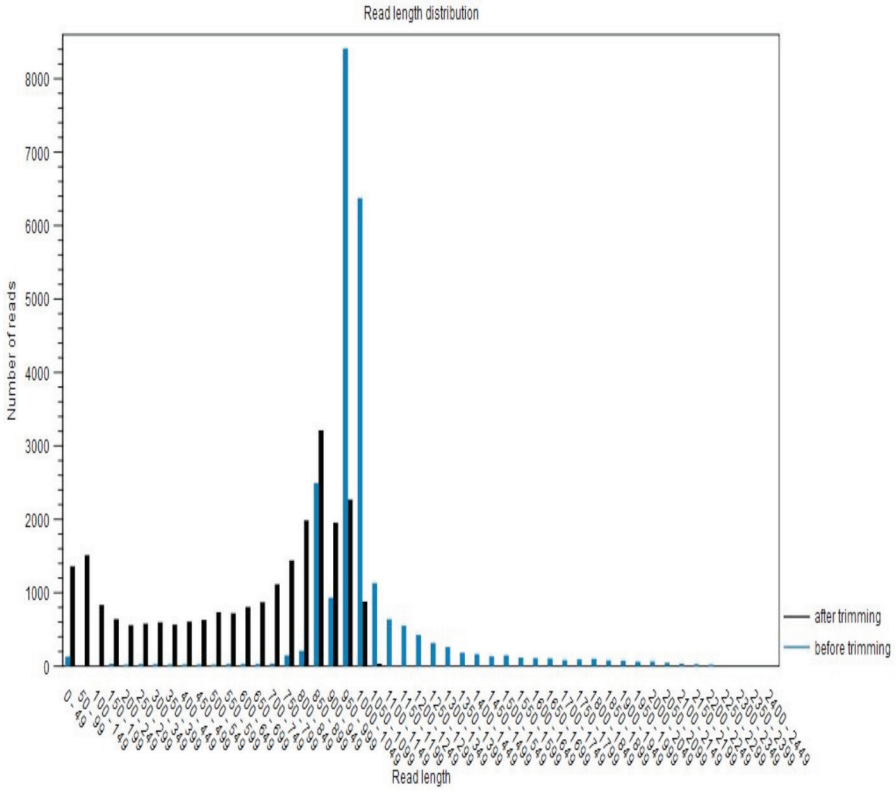


Figure 37 –*Representation of the Library 1 read length before and after performing the chosen trimming (graphics produced by CLC genomics software).*

III.3.1 Results of the *de novo* assembly of *D. gigas* genome using CLC genomics software

The next two tables (Table 15 and 16) present the results of the *de novo* assembly, where contig assembly was done by using 23,959 high quality Sanger reads, that resulted in 967 contigs by the CLC method. The main highlighted results are the 21% reads that could not be integrated in any contig and the very low percentage of genome covered (only 44%), obtained with CLC Genomics Workbench.

Table 15 – *D. gigas* genome general outcome obtained after performing the *de novo* assembly using the raw data from Sanger Sequencing

Total number of Reads	Average size of the reads	Total number of bases	Number of contigs after assembly	% un-mapped reads	% of the genome covered	Medium coverage	Total size
11,074	663	7,349,107	967	21.33%	44%	1.59	1,697,742

The next table summarizes the calculated lengths, in bp, of the biggest contigs until 75% (N75), 50% (N50) and 25% (N25) respectively are reached, being the represented value indicative of the smallest contig used to reach the respective percentage. Minimum, maximum and average contig size is also represented. As can be seen, the size of the obtained contigs is very small, being the average contig length of 1.7Kb (when the reads, per se, are 600 bp after trimming). The maximum length contig obtained by our assembling method, out of an initial number of >22,000 high quality reads, was 9.5 Kb. No wonder that only 44% of the genome was covered. On average, the contig length was very low, indicating that the Sanger raw data was highly insufficient despite the almost 25.000 sequencing reactions produced.

Table 16 – Contig measurements relevant data. N75, N50 and N25.

Contig Measurements	
N75	1,372
N50	1,864
N25	2,733
Minimum	228
Maximum	9,578
Average	1,756

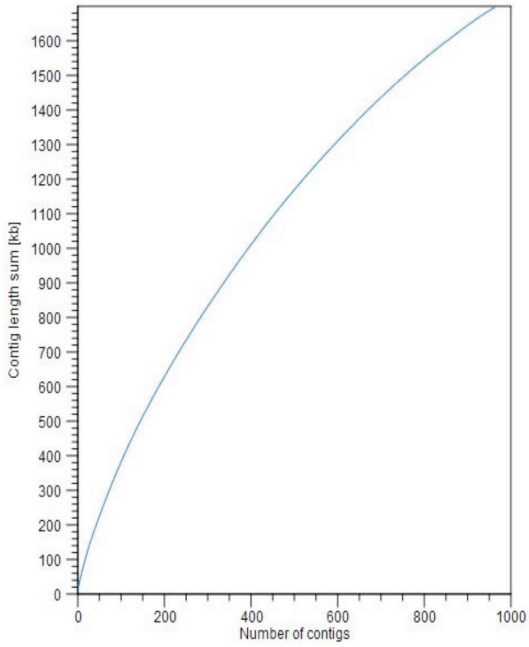


Figure 38 – *Graphical representation of the accumulated contig length.*

Fig.38 summarizes the contig length, represented in the y-axis and the number of contigs, represented in the x-axis. This accumulated contig length curve is obtained from the analysis of the *de novo* assembly. As mentioned above, the *de novo* assembly only yielded coverage of 44% of the whole *D. gigas* genome, this is also visible in this figure where an accumulated contig length of approximately 1700 kb from the total is obtained with 967 contigs, contrasting with the 3700 kb of the complete genome. The shape of the graphic (almost linear) is indicative of a less successful *de novo* assembly (ideally, the line should fit entirely up to 200 contigs in the x-axis)

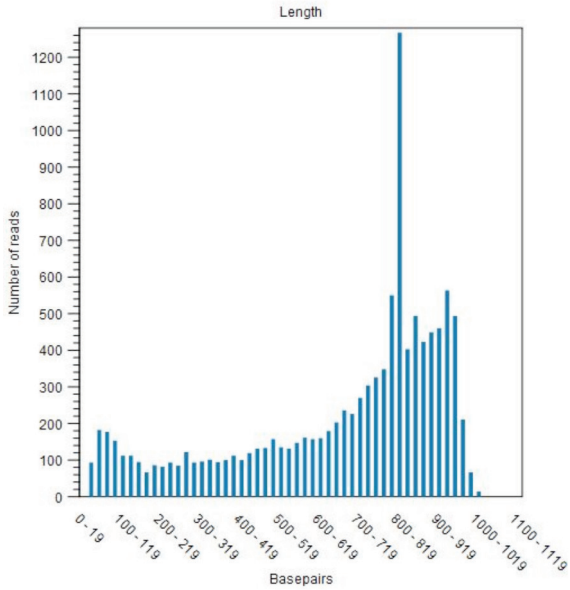


Figure 39 – *Distribution of the matched read length.*

Fig.39 shows that sequences with length ranging from approximately 780 bp to 920 bp are visibly the more matched than lower, which is in accordance with the previously represented distribution of read length since the number of reads with length over 700 bp was higher, and the 800-819 bp peak was also present.

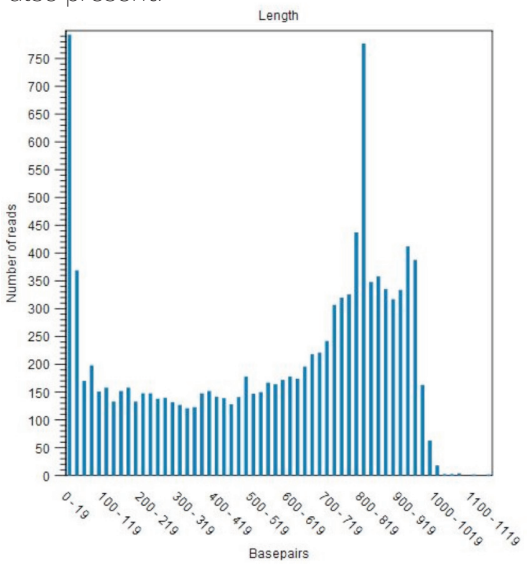


Figure 40 - *Distribution of the unmatched read length.*

Fig.40 shows that there were many unmatched reads to the genome of *D. gigas*, sequences with length ranging from approximately 780 bp to 920 bp and sequences with length under 40 bp are visibly the more unmatched than lower. It should be noted that in the total read distribution, the length over 700 bp was also higher, and the 800-819 bp peak was also present.

III.3.2.- Results obtained by mapping the high quality Sanger reads against the *D. gigas* genome

This exercise of re-sequencing *a posteriori* using our raw data and map it against the reference genome published by ourselves (Morais-Silva *et al*, 2014), gives us a very good idea of the quality of the original reads. Mostly, after mapping our Sanger raw data covers 85% of the genome and 65% of the plasmidic DNA, allowing the conclusion that, either the plasmidic DNA is under represented in the cloning process for the library construction, or its fragments are less viable (e.g. toxic) for the hosting *E.coli* bacteria used for the library construction.

Table 17 – General outcome obtained after the mapping of the high quality reads vs the reference genome.

	% of mapped reads	% of unmapped reads	% of genome covered	Average coverage
chromosome	62,26%	37,74%	85%	2.75
Plasmid	0.88%	99,12%	65%	1.40

As described in the tasks, for the mapping we excluded all the reads that have more than 20% difference when comparing to the reference genome, and this results in the exclusion of, roughly, 38% of the reads. But even if this high number of reads is excluded, the remaining still map in 85% of the reference genome, at an average cover per base of only 2.75 times. This allows the conclusion that the libraries had other fragments besides *D. gigas* DNA, but that the right clones had extremely high quality and representation all over the genome.

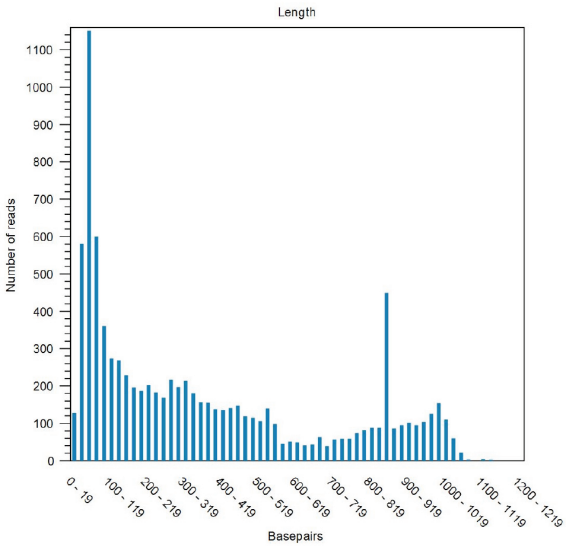


Figure 41 – *Distribution of the unmapped read length*

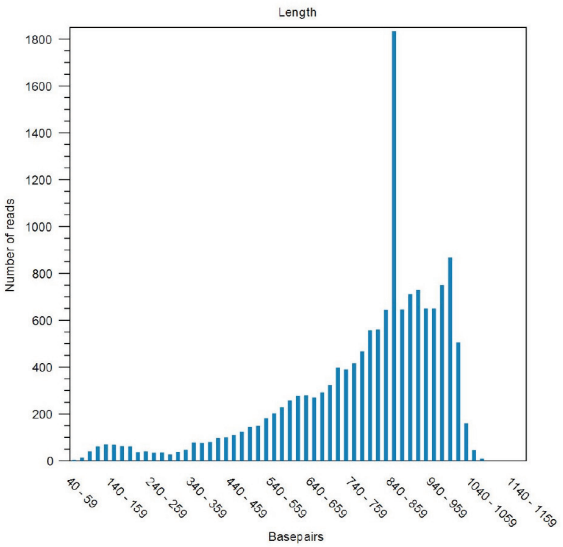


Figure 42 – *Distribution of the Mapped read length*

Figs 41 and 42 also show us that the reads that are mostly unmatched are the ones with smaller length, while the mapped reads are mostly the reads with lengths between 750 and 900 bp, allowing the conclusion that smaller reads have less quality and, proportionally higher concentration of the DNA non-specific constructs.

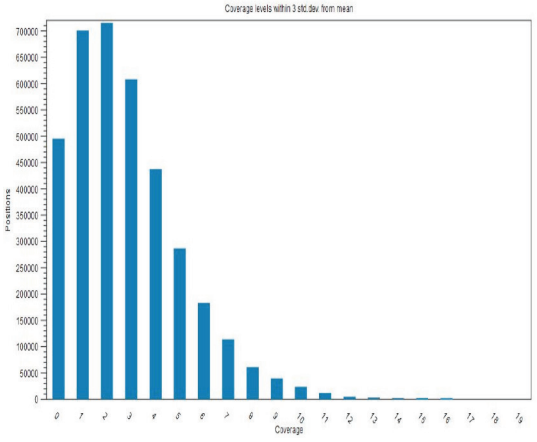


Figure 43 – *Distribution of the coverage within 3 standard deviation from the mean obtained after the mapping to reference.*

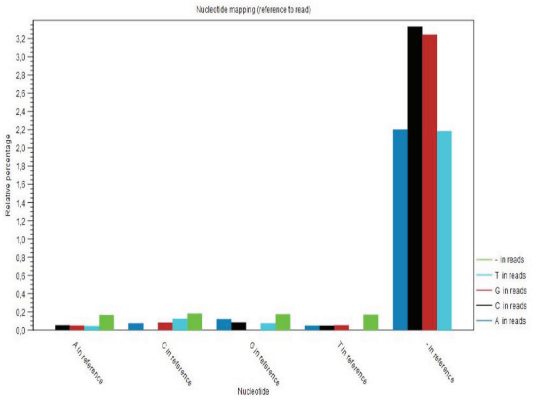


Figure 44 - *Nucleotide mapping relative errors – Representation of the most often found substitutions for each type of base or gap in the reference sequence.*

Fig.43 shows that most of the covered positions have a 1 – 4 coverage. Approximately 500,000 positions (495,388) have 0 coverage. 2,889 positions have coverage above 19 (not shown in the graph).

Fig 44 shows that Cs and Gs are more often substituted by a gap in the reference sequence (with more than 1% difference when compared with As and Ts). This is in accordance with the increasing error rate for genomes with high CG content. It should be noted that only mismatches are plotted, the matches are not included. The rate of mistaken insertions in the individual reads was extremely high, comparatively, especially the insertion of C's and G's.

As a conclusion, we can say that the Sanger raw data was insufficient, and mostly concentrating in specific zones where the library method produced more clones.

III.4 - Analysing the DNA sequencing raw data obtained by the 454 platform. *De novo a*ssembly

In order to be able to assemble the data obtained, the CLC Genomics Workbench was used the two major experiments performed in the 454 Pyrosequencing platform produced 26,5 and 46 Megabases of sequence data from genomic DNA of *D. gigas*.

The results obtained can be summarized as follows:

Table 18 – Results obtained after the trimming of the raw data.

Raw data & trimming			
	Biocant (subcontractor)	Keygene (subcontractor)	Taken together
Number of reads	275,549	459,801	735,350
Average read length (bp)	96.7	96.2	96.4
Average trimmed read length (bp)			93.88
Number of not trimmed reads			489,395
Number of reads after trimming			735,350
Successful (%)			100%
Total Mbases	26.5	46	73

The raw data obtained in Biocant had less quality and fewer number of reads inserted into contigs. The raw data from Biocant, if isolated, gave a huge number of 4,421 contigs after assembly, while the Keygene run, *per se*, gave "only" 1,124 contigs, just a bit more than the whole raw data from Sanger. Taken together the number of contigs lowered to 568, with an average contig size of 6.5 Kb.

Table 19 – Results of *de novo* assembly obtained from the two individual sequencing runs on 454 platform, performed by the subcontracted companies, Biocant and keygene.

Outcome of the 2 raw data from 454 runs			
	Biocant	Keygene	Taken together
"Reads" insert in	267,556	449 904	705,439
Number of "contigs"	4 421	1 124	568
Number of bases	3 628 162	3,742 552	3,699,903

Table 20 – Assembly performed with the CLC bio relevant data.

Contig Measurements	
N75	6,600
N50	12,177
N25	18,645
Minimum	207
Maximum	45,814
Average	6,461
Number of contigs	568
Number of bases	3,699,903

In table 20 it is summarized the calculated lengths of the biggest contigs for covering 75% of the genome, 50% and 25% of the complete genome, respectively, are reached, being the represented value indicative of the smallest contig used to reach the respective percentage. Minimum, maximum and average contig size is also represented. Finally the number of contigs and number of bases obtained after the assembling process. The biggest contig after assembly had a size of 46 Kb approximately.

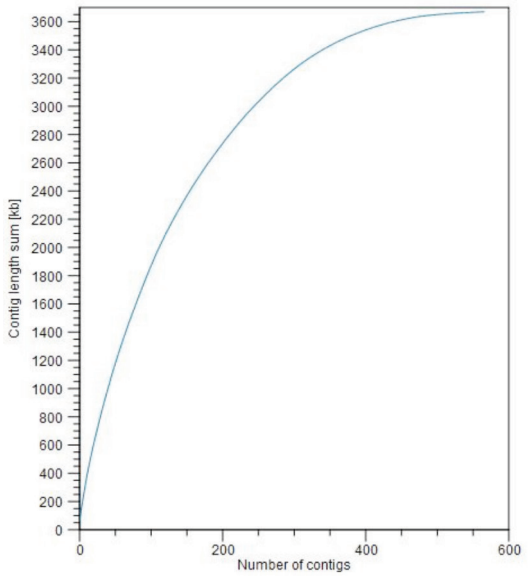


Figure 45 - *Graphical representation of the accumulated contig length.*

Figure 45 shows that the *de novo* assembly yielded a coverage of 99% of the whole *D. gigas* genome and shows that most of the assembly of the genome needs "only" 400 contigs. Summarized contig length is represented in the y-axis and the number of contigs is represented in the x-axis. This accumulated contig length curve is obtained from the analysis of the *de novo* assembly.

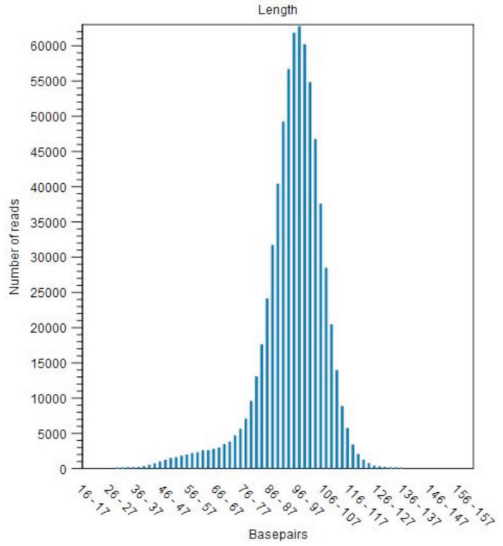


Figure 46 - *Distribution of the matched read length.*

The figure 46 shows that sequences with length ranging from approximately 80 bp to 105 bp are visibly the more matched than lower, which is in accordance with the previously represented distribution of read length.

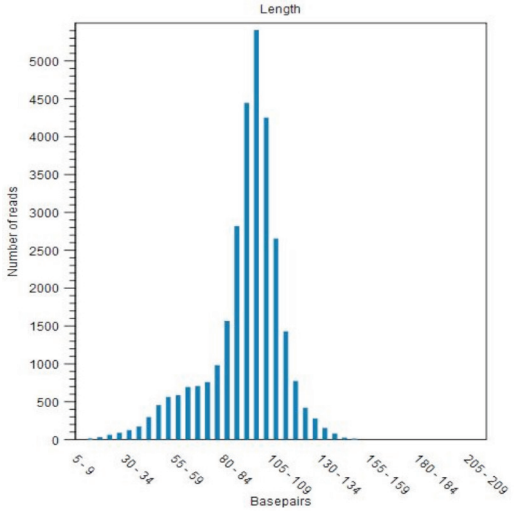


Figure 47 - *Distribution of the unmatched read length –*

The figure 47 shows that, in proportion, the most unmatched reads are the lowest length ones, suggesting that these short reads are in higher proportion and are the result of errors created by the sequencing method.

III.4.1 Results obtained by mapping the 454 contigs against the *D. gigas* reference genome

In order to evaluate the quality of the reads from the raw data obtained by the 454 method, we decided to map, *a posteriori*, this raw data against the completed sequence of *D. gigas*, deposited by us at NCBI. The main findings were that only 3.5% of the raw data did not match the chromosome, out of which 2.21% matched the plasmid and, with 454 data, we covered 96% of the genome, 17 times on average each base.

This led to the conclusion that 4% of the genome was not represented in the 454 library. This could be related to the DNA characteristics of this non-represented portion of the genome.

Table 21– General outcome of the CLCbio ***de novo*** assembly, from the two 454 raw data taken together.

	% of mapped reads	% un-mapped reads	% of the genome covered	Average coverage
chromosome	96.5%	3.50%	96%	17.27
plasmid	2.21%	97.79%	96%	14.96

This allows the conclusion that the library preparation method had extremely high quality and representation. Moreover since 1.29% of the reads did not match either in the plasmid or in the chromosome, one can infer that error rate associated with the method is no bigger than 1.29%.

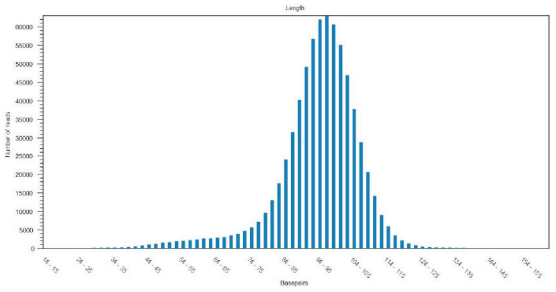


Figure 48 – *Distribution of the mapped read length.*

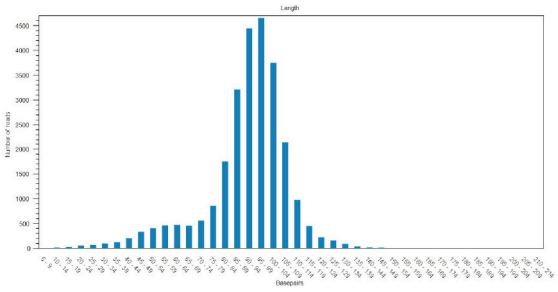


Figure 49 – *Distribution of the unmapped read length.*

The figs 48 and 49 also show us that the reads that are mostly unmatched are the ones with smaller length, while the mapped reads are, in proportion, the reads with 100bp. This allows the conclusion that smaller reads have less quality, and are, probably, generated dummy reads from the reading method or mistakes of the library preparation.

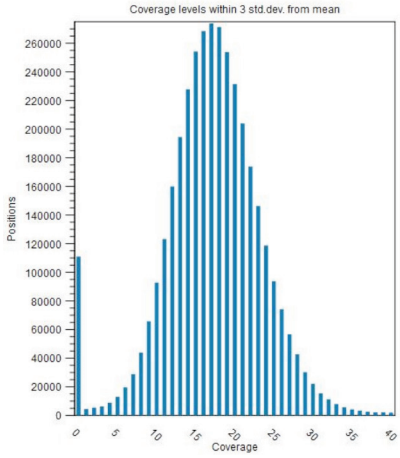


Figure 50 - *Distribution of the coverage within 3 standard deviation from the mean obtained after the mapping to reference.*

The fig 50 shows that most of the covered positions have a 10 to 25 coverage. Approximately 110,000 positions had 0 coverage corresponding to the gaps between the contigs. While 17,589 positions had coverage above 40 (not shown in the graph). This means that 110Kb of the genome could not be represented in the library, and that, on average, each unfilled gap had a mean size of 200bp between contigs.

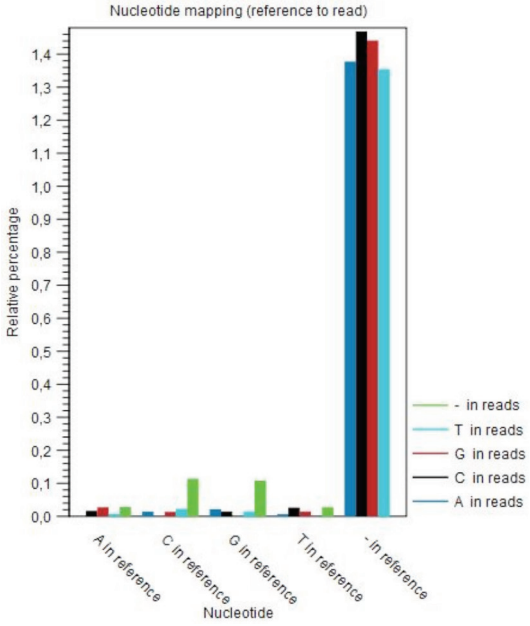


Figure 51 - *Nucleotide mapping relative errors – Representation of the most often found substitutions for each type of base or gap in the reference sequence.*

The fig 51 shows Cs and Gs are more often substituted by a gap in the reference sequence than. As and Ts (the difference is very slight, around 0.1%). It should be noted that only mismatches are plotted, the matches are not included. Comparing to Sanger reads, this 454 method deletes more C's and G's, most probably in homopolymeric regions. On opposite, the mistaken inserted bases happen with less frequency and at the same rate for all nucleotides.

III.5 - Analysing the raw data from *D. gigas* sequencing on the Illumina platform for the *de novo* assembly of its genome

On overall using the Illumina platform, by subcontracting Washington University and Baseclear, we got more than 20 million reads of 51 bp on average, making a total of 1 Gbases of raw data. For trimming this data, since the quality of the reads (Phred score) was overall high in the raw data, the parameters defined were allowing only 3 ambiguous nucleotides and a maximum of 10 bases with Phred score under 20 with this 1bp of raw data, de novo assembly results in only 130 contigs (tables 22 and 23).

Table 22 - Results obtained after the trimming of the raw data

Raw data & trimming	
Number of reads	20,167,690
Average read length (bp)	51.0
Average trimmed read length (bp)	50.9
Number of not trimmed reads	19,767,012
Number of reads after trimming	20,167,690
Successful (%)	100%
Total Megabases	1,012

Table 23 – Assembly performed with the CLCbio relevant data.

Contig Measurements	
N75	34,172
N50	60,650
N25	90,262
Minimum	1,032
Maximum	199,432
Average	28,918
Number of contigs	130
Number of bases	3,759, 314

Table 23 summarizes the calculated lengths of the biggest contigs until 75%, 50% and 25% respectively are reached, being the represented value as an indication of the smallest contig used to reach the respective percentage. Minimum, maximum and average contig size is also represented. Finally the number of contigs and number of bases obtained after the assembling process. The biggest contig after assembly had a size of approximately 200 Kb. Up to 75% of full coverage of the genome was reached, where the smallest contig had a size of 35 Kb approximatively.

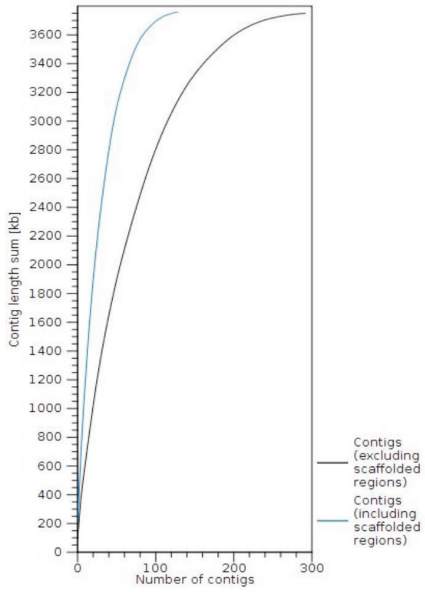


Figure 52 - *Graphical representation of the accumulated contig lengths including and excluding scaffold regions*,

The fig 52 shows a summarized representation of contig length in the y-axis and the number of contigs in the x-axis. This accumulated contig length curves are obtained from the analysis of the *de novo* assembly. The blue line represents the accumulated contig length when performing scaffolding and the black line represents the accumulated contig length without the scaffolding step. The shape of the line is indicative of a successful *de novo* assembly, especially when using the scaffolding option

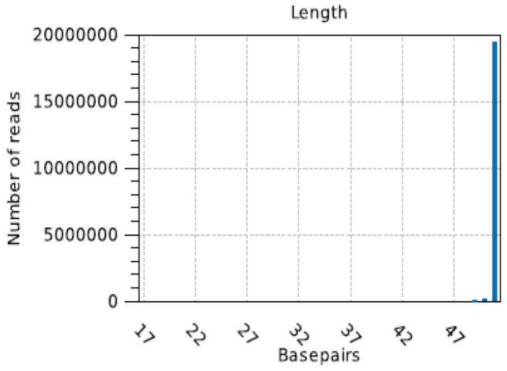


Figure 53 - *Distribution of the matched read length - as visible the matched read length distribution is similar to the previously represented distribution of read length*

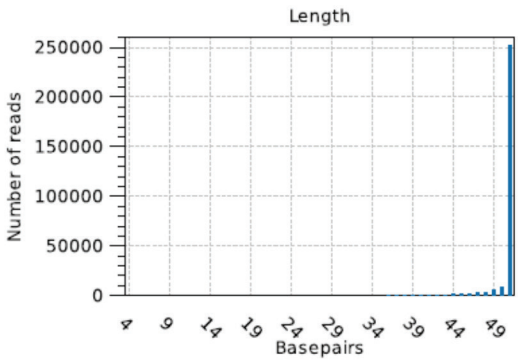


Figure 54 - *Distribution of the unmatched read length*

Figures 53 and 54 show that most of the assembly uses 50 bp reads, which are in great majority. The 250,000 reads out of 20 million reads were discarded and were not used in the *de novo* assembly.

III.5.1 Results obtained by mapping the Illumina raw data against the *D. gigas* reference genome

The *a posteriori* mapping of the reads against the genome gave an average coverage of 260-fold, 98% of the whole genome covered and 3.76% of reads unmapped. This is an evidence of an error rate of a maximum of 3.5% of the method, but also a good evidence of a widespread representation of most of the genome in the produced DNA library.

Table 24 – General outcome of the mapping to reference.

	% mapped reads	% un-mapped reads	% of the genome covered	Medium coverage
chromosome	96.3%	3.76%	98%	263.4
plasmid	2,66%	97,34%	96%	264.3

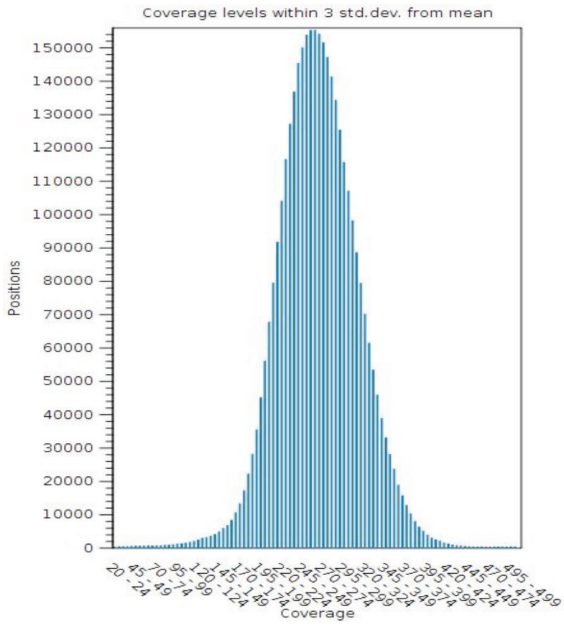


Figure 55 - *Distribution of the coverage within 3 standard deviation from the mean obtained after the mapping to reference.*

Fig 55 displays the findings that, out of 3,6 Kb of the genome, 3,578,999 positions have coverage between 21 and 505; 85,800 positions have coverage below 21 (not shown in graph) and 29,200 positions have coverage above 505 (not shown in graph). In conclusion, 85 Kb of the genome is less properly represented in the library, while 30 Kb of the DNA was super represented and highly represented in the sequencing, which could represent the sum of all repetitive elements throughout the genomic DNA.

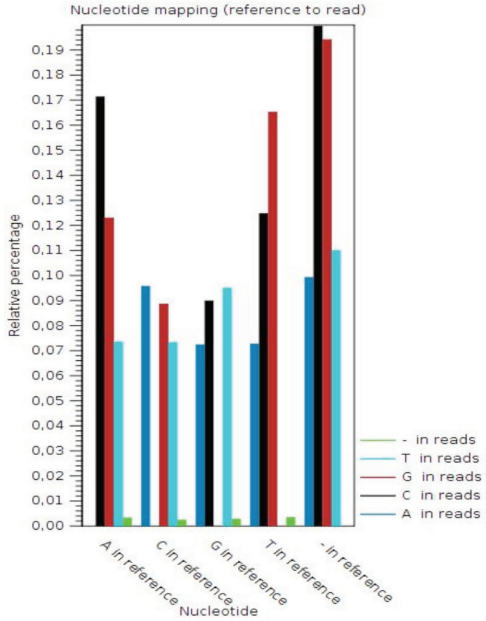


Figure 56 - *Nucleotide mapping relative error count*

Fig 56 shows a representation of the mismatches for each type of base or indels when comparing to the reference sequence. In the figure it's visible that the overall error rate is very low, Mistaken insertions occur at a low frequency rate but still higher in proportion for Cs and Gs. Deletions are very rare and mismatches occur more frequently for A's and T's than for C's and G's. It should be noted that only mismatches are plotted, the matches are not included.

III.6 - Analysing the sequencing raw data obtained with the Ion torrent platform - *de novo* assembly

The raw data from this experiment consisted of 2,085,311 reads, with an average size of 175 bp. Overall, the experiment gave an additional 360 Mbases to help completing the *D. gigas* genome. These results are illustrated in Table 25 and Fig. 55.

The trimming parameters are important to the subsequent steps, the quality of the reads influences the mapping and assembly steps. The trimming parameters defined allowed a maximum of 3 ambiguous nucleotides and a subsequent trim that removed sequences with less than 12 nucleotides. After trimming, the higher quality raw data consisted of 1,977,262 reads with an average size of 163 bp, making it a total of approximately 322 Mbase of high quality DNA sequences.

Table 25 - Results obtained after the trimming of the raw data obtained on the PGM machine (Ion Torrent technology) at STAB VIDA from ***D. gigas*** DNA.

Raw data & trimming from Ion Torrent run	
Average read length (bp)	174
Average trimmed read length (bp)	163
Number of not trimmed reads	624,942
Number of reads before trimming	2,085,311
Number of reads after trimming	1,977,262
Successful (%)	94.84%

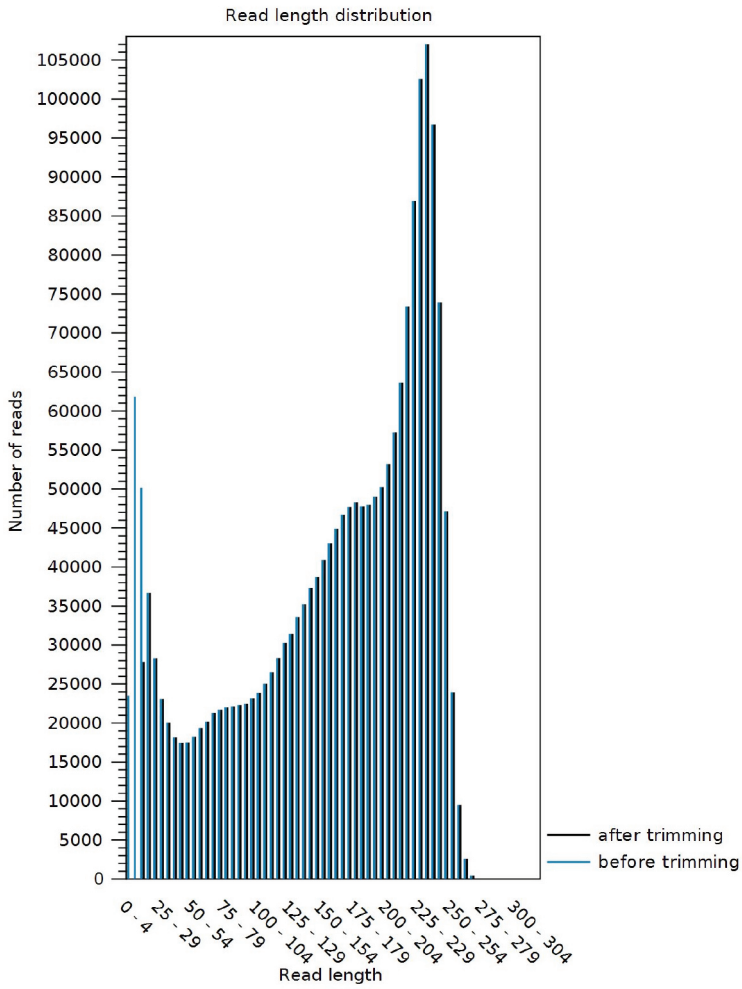


Figure 57 - *Distribution of the read length before and after the trimming.*

The *de novo* assembling of the Ion Torrent raw data after trimming resulted in a total of 923 contigs, with an average size of 3.6 Kb each, and an average coverage per base of 75 times. 15% of the genome was not covered. These numbers are shown in the next table. These results, when compared to previous experiments, are in the same order of number of contigs but with higher coverage. The lengths of the biggest contigs to reach 75%, 50% and 25% respectively are represented in Table 26. The biggest contig had a size of 38 Kb, while the minimum contig had a size of 488 bp.

Table 26 - Resumed outcome of the treated raw data obtained from the Ion Torrent run.

Contig Measurements	
N75	3,316
N50	6,530
N25	10,831
Minimum	488
Maximum	37,864
Average	3,624
Reads insert in contigs	1,707,403
Number of contigs	923
Number of bases	3,344,694

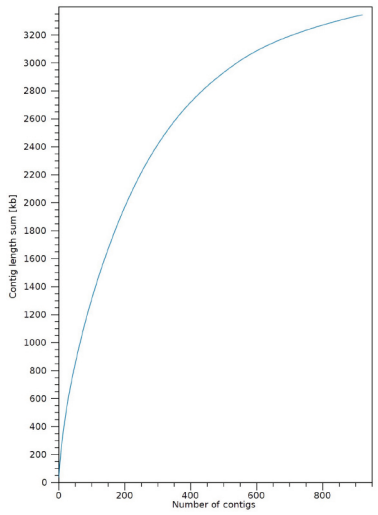


Figure 58 - *Graphical representation of the accumulated contig length*

The fig 58 represents the accumulated contig length in the y-axis and the number of contigs is represented in the x-axis. This accumulated contig length curves are obtained from the analysis of the *de novo* assembly. Since the line is closer to the y-axis rather than its linear counterpart, the figure is indicative of good quality of raw data and *de novo* assembly.

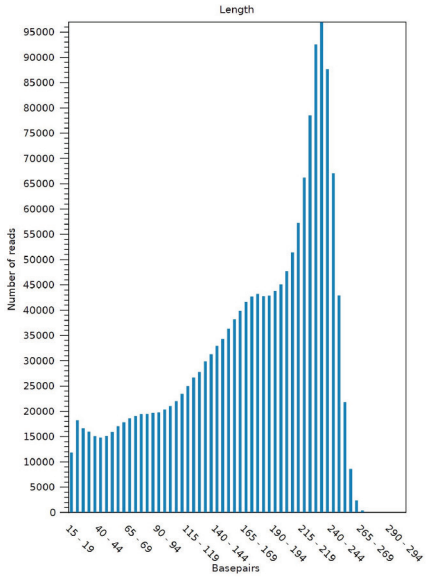


Figure 59 - *Distribution of the matched read length*

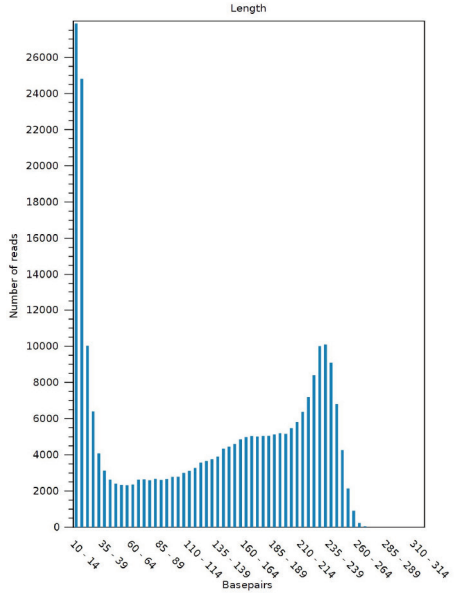


Figure 60 - *Distribution of the unmatched read length*

As visible in Fig 59, the read length of the matched sequences is widely distributed, from approximately 15 bp to 270 bp, even if sequences ranging from 200 to 240 bp are visibly more matched than others, which is in accordance with the previously represented distribution of read length. Fig 60 shows that the smaller sequences, ranging from 10 to 20 have an obvious unmatching tendency.

III.6.1 Results obtained by mapping reads against the *D. gigas* reference genome

By mapping the Ion Torrent reads, a total of 1.9 million reads with average size of 170 bp, we concluded that 3.7% of these reads do not match with any sequence of the reference genome, and that only 85% of the genome is covered, even if at a coverage average of 75 times. This gives us some clues to the error rate of the technology that we conclude to be in the order of 1.79% and not 1% as claimed by the supplier, at least with high GC .genomes. Moreover, if only 85% of the reads match with the genome, it means that the library construction missed 15% of the DNA. This is also supported by Figs 59 and 60.

Table 28 - General outcome of the mapping to reference.

	% mapped reads	% un - mapped reads	% of the genome covered	Average coverage
chromosome	96,44	3,66%	85%	75,3
plasmid	1,87%	98,13%	96%	61,8

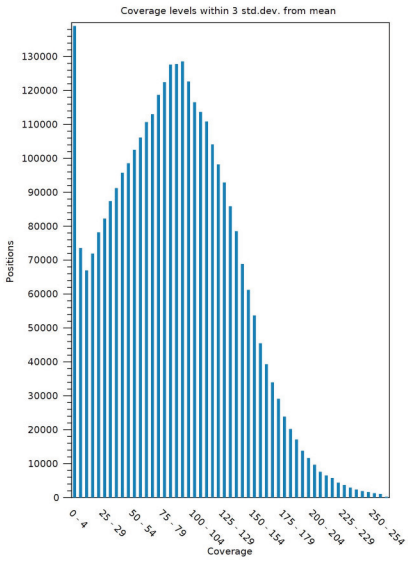


Figure 61 - *Distribution of the coverage within 3 standard deviation from the mean obtained after the mapping to reference.*

The fig 61 shows that most of the positions of the DNA sequence have an average coverage in between 75 and 125 times. 3.334.008 positions have coverage between 1 and 260 and 10.686 positions have coverage above 260 (not shown in graph).

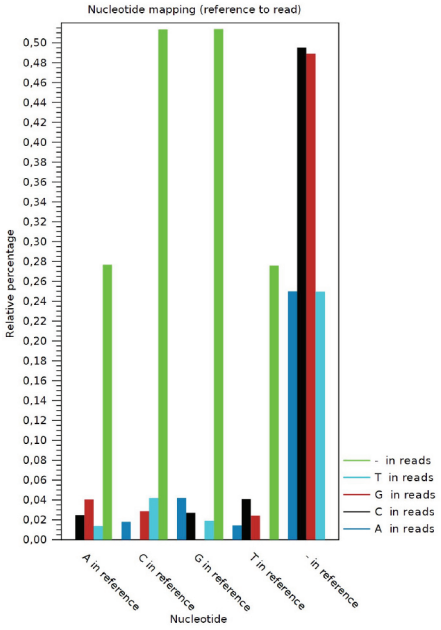


Figure 62 - *Nucleotide mapping relative errors - Representation of the most often found substitutions for each type of base or gap in the reference sequence.*

In the Fig 62 it is visible that the rate of mismatches was very low for all bases, but for all bases the rate of deletions was proportionally a lot higher than the raw data of for instance Illumina. Mistaken insertions rate is low but now as low as in Illumina, and again especially the Cs and Gs seem to be more prone to be inserted in the reads than As and Ts. It should be noted that only mismatches are plotted but the matches are not included.

III.7 – Comparison of the four methods: Some assumptions taken after our in-house bioinformatics analysis

The comparison of the obtained raw data from the four different methods, before and after trimming, is summarized in table 29. It is also shown the cost of each Megabase (Mb) of DNA Sequence, although those costs have to be referred to the date of the experiment (today, the costs are much lower already).

Table 29 - Summary of all raw data

	Sanger	454	Illumina	Ion Torrent	Taken together
Total number of bases before trimming (Mb)	24.8	74	1,012	362.8	1,474
Total number of bases after trimming (Mb)	13.6	73	1,012	322.3	1,421
Average read length (bp)	1000	96	51.0	174	NA
Average trimme d read length (bp)	603	94	50.9	163	NA
Number of reads before trimming	23,959	735,350	20,167,690	2,085,311	23,012,310
Number of reads after trimming	22,536	735,350	20,167,690	1,977,262	22,902,838
Successful %	87%	100%	100	94.84%	NA
Total cost of the experiment (€)	118,600	40,732	6,010	1,900	167,342
Cost per Mb (€)	4,780	550	6	5	118

Back in 2008-2009, STABVIDA paid more than 40,000€ to obtain 70 Megabases of raw data of the genome of *D. gigas* with little or none bioinformatics included. At present, STABVIDA offers the clients a commercially full in-house service of generating 800 Megabases of data with bioinformatics included for 800€ (meaning 1€ for each megabase of sequencing data).

The comparison of the results of the *de novo* assembly are summarized in table 30.

Table 30 - Summary and comparison of *de novo* assembly, individually and taken together.

	Sanger	454	Illumina	Ion Torrent	Taken together
N75 (bp)	1,372	6,600	34,172	3,316	56,223
N50 (bp)	1,864	12,177	60,650	6,530	79,621
N25 (bp)	2,733	18,645	90,262	10,831	99,172
Minimum (bp)	228	207	1032	488	33,303
Maximum (bp)	9,578	45,814	199,432	37,861	212,271
Average (bp)	1,756	6,641	28,918	3,624	70,474
Number of contigs	967	568	130	923	45
Number of bases (Mb)	7.4	3,699	3,759	3,344	13,171,5
Number of reads	11,074	705,439	19,767,012	1,977,262	19,059,272
% un-mapped reads	21.3	1.2	3.76	3.66	16.7
% of the genome covered	44	96	98	85	98

Comparing the results obtained by mapping, *a posteriori*, the reads against the reference complete *D. gigas* genome is shown in table 31 for the bacterial chromosome and in table 32 for the bacterial plasmid.

Table 31 - Chromosome of reference genome used for mapping the raw data

	Sanger	454	Illumina	Ion Torrent	Taken together
% mapped reads	62	96.5	96.3	96.4	97.60
% un-mapped reads	37	3.5	3.7	3.6	2.40
% of the genome covered	44	96	98	85	98
Average coverage	2.75	17	263	75.3	370.8

Table 32 - Plasmid of reference *D.gigas* genome used for mapping all raw data, individually and taken together.

	Sanger	454	Illumina	Ion Torrent	Taken together
% mapped reads	0.88	2.2	2.7	1.87	2.73
% un-mapped reads	99	97.7	97.3	98.1	97.27
% of the genome covered	65	96	96	96	96
Average coverage	1.4	15	264	62	350.5

Chapter IV – The complete genome of *D. gigas* and its annotation

Preface

In this chapter it is shown the result of the full assembling and the annotation of the *Desulfovibrio gigas* genome.

All raw data obtained from the four different methods was assembled by the professional team led by Jeronimo Ruiz from FioCruz, BH, Brazil, using Velvet software and the consensus genomic sequence was obtained with Phrap. Protein-coding sequences were manually curated essentially by the *D. gigas* genome team at ITQB and STAB VIDA using the Artemis Software. The main results of this assembling and annotation exercise can be seen in tables 1 and 2, respectively. These findings were submitted for publication in March 2014 to "Open Microbiology" and are now published. The data of the publication was 15th May of 2014. A total of 3,370 genes were found in the bacterial chromosome, plus one in the plasmid. Of these, 3,273 genes are protein-coding, but it was not assigned the function to 999 of them.

Acknowledgments: The author's contribution to this chapter was a modest participation in the genome's annotation exercise, that took several months, and all the previous work concerning the MOPP gene and MOPP operon. The author acknowledges Fábio Silva, Catarina Pimentel, Cátia Santos and Jeronimo Ruiz for the professional work in assembling and annotating the genome as well as Ulrich Thoenes for his contribution to the cloning and sequencing of the MOPP gene.

"*If you're walking down the right path and you're willing to keep walking, eventually you'll make progress*"

Barack Obama

IV.1 – The submission of the complete genome sequence to NCBI in 2013 and the publication in 2014

This chapter starts by the recent publication Morais-Silva *et al*, where our team, present the major findings from the genomic sequence of *D. gigas*. The following tables 33 and 34 summarize the general genome features of *Desulfovibrio gigas*.

Table 33 – General genome features of *Desulfovibrio gigas*

Features	Value	% of total
Genome		
Genome Size (bp)	3,693,899	100
DNA coding region (bp)	3,249,714	87.98
G+C content (bp)	2,341,530	63.39
Extrachromossomal elements	1	
Number of replicons	1	
Total number of genes	3370	100
Stable rRNAs		
rRNAs	3	0.09
tRNAs	48	1.42
Protein-coding genes	3273	97.09
Genes density (bp/gene)	1128	
Average length of a gene (bp)	993	
Pseudogenes	47	1.39
Genes with as signed COG	2273	67.45
Selenocysteine-containing proteins	9	
Genes without assigned function	999	29.64
Poorly characterized genes	395	11-72
Other elements		
CRISPR repeats	6	
Cas operons	2	
Transposases	17	
Mobile elements	1	

Table 34 - *Desulfovibrio gigas* gene classification by pathway

Pathway	Number of genes
Central Metabolism	
Alcohol metabolism	6
ATP synthesis	17
Beta oxidation	5
Embden-Meyerhof-Parnas Pathway	23
Entner-Doudoroff Pathway	3
Pentose phosphate Pathway	8
TCA Cycle	6
WoodLjungahl Pathway	8
Glyoxylate Cycle	1
Oxidation of pyruvate to acetyl-CoA and acetate	20
Lactate metabolism	6
Beta Lactamase proteins	8
Formate Metabolism	7
Fumarate Metabolism	6
Methylglyoxylate Cycle	3
General metabolism	
Sulfate Metabolism	22
Nitrogen Metabolism	31
Transcriptional Factors Sigma 54	13
Response to Oxygen	17
Energy Conservation	
Membranar energy complexes	65
Hydrogenases	15
Cytochromes	16
Hdr-like proteins	12
Nfn complexes	5
Selenocystein-containing proteins	9
Miscellaneous	
CRISPR proteins	10
Chemotaxis proteins	84

The genome of *D. gigas* (CP006585) consists of one circular chromosome of 3,693,899 base-pairs (bp) having 3370 genes of which 3273 are protein-coding. The genome has a G+C content of 63.4% that reflects a biased codon usage. Indeed, *D. gigas* prefers high G+C codons (66.87%), with a clear preference for cytosine (C) in the 3rd position (82.03%). The genome is very compact as observed by its gene density of 1128 bp per gene and the average length of each gene is 993 bp. It contains 17 transposases, whereas in other SRB genomes this number is in average 34 (Bennett 2004). This relative low number of transposable elements in D. gigas may indicate a low rate of reorganization of its genome. Other features include 47 pseudogenes and 48 tRNAs (Table 1), as well as 9 selenocysteine containing proteins. Surprisingly, one single operon of rRNA was found in *D. gigas* in contrast to what was detected in other *Desulfovibrio* spp. that contain between 3 and 6 operons. The plasmid of this bacterium (CP006586) has a size of 101,949 bp, containing 75 ORFs, of which 72 are coding regions.

Table 35 - General plasmid features of *Desulvobrio gigas*.

Features	Value
Size (bp)	101,949
G + C contente (bp)	64,081
DNA coding region (bp)	79,425
Pseudogenes	3
Protein-coding genes	72
Gene density (bp/gene)	1415
Average length of a gene (bp)	1103

The sequencing of the *D. gigas* genome provides insights into the integrated network of energy conserving complexes and structures present in this bacterium. Comparison with genomes of other *Desulfovibrio spp.* reveals the presence of two different CRISPR/Cas systems in *D. gigas*. Phylogenetic analysis using conserved protein sequences (encoded by rpoB and gyrB) indicates two main groups of *Desulfovibrio* spp, being *D. gigas* more closely related to *D. vulgaris* and *D. desulfuricans* strains.

MicrobiologyOpen

Open Access

ORIGINAL RESEARCH

Genome sequence of the model sulfate reducer *Desulfovibrio gigas*: a comparative analysis within the *Desulfovibrio* genus*

Fabio O. Morais-Silva[1,a], Antonio Mauro Rezende[2,a], Catarina Pimentel[1], Catia I. Santos[1], Carla Clemente[3], Ana Varela–Raposo[1], Daniela M. Resende[2], Sofia M. da Silva[1], Luciana Márcia de Oliveira[2,4], Marcia Matos[3], Daniela A. Costa[2], Orfeu Flores[3], Jerónimo C. Ruiz[2] & Claudina Rodrigues-Pousada[1]

[1]Instituto de Tecnologia Quómica e Biológica – Antonio Xavier, Universidade Nova de Lisboa (ITQB-UNL), Av. da República – Estação Agronómica Nacional, 2780-157, Oeiras, Portugal
[2]Grupo Informática de Biossistemas, Centro de Pesquisa René Rachou – FIOCRUZ, Belo Horizonte, Minas Gerais, Brazil
[3]STAB VIDA - Madan Parque Rua dos Inventores s/sala 2.18, 2825-182, Caparica, Portugal
[4]Departamento de Microbiologia, Programa de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais, Brazil.

Keywords
Analysis, comparative genomics, *Desulfovibrio gigas*, genome.

Correspondence
Claudina Rodrigues-Pousada, Instituto de Tecnologia Química e Biológica António Xavier, Av da República (EAN), 2780-157 Oeiras, Portugal. Tel: +351214469624; Fax: +351214469625;
E-mail: claudina@itqb.unl.pt

Jerónimo C. Cruz, Grupo Informática de Biossistemas, Centro de Pesquisa René Rachou – FIOCRUZ, Belo Horizonte, 3019002 Minas Gerais, Brasil. Tel: +55 31 3349 7700; Fax: +351 21 043 860;
E-mail: jeronimo@cpqrr.fiocruz.br

Orfeu Flores, STAB VIDA - Madan Parque Rua dos Inventores s/sala 2.18, 2825-182 Caparica, Portugal. Tel: +351960022300; Fax: +553132953115;
E-mail: orfeu@stabvida.com

Present address
Antonio Mauro Rezende, Departamento de Microbiologia, Centro de Pesquisas Aggeu Magalhães – FIOCRUZ PE, Av. Professor Moraes Rego, 50670-420,Recife/PE, Brazil

Funding Information
This work was supported by Fundação para Ciência e Tecnologia FCT through grants PTDC/BIA-IC/104030/2008 given to C.R.P., Pest-OE/EQB/LA0004/2011 given to ITQB. Agência de Inovação (ADI) also supported

Abstract

Desulfovibrio gigas is a model organism of sulfate-reducing bacteria of which energy metabolism and stress response have been extensively studied. The complete genomic context of this organism was however, not yet available. The sequencing of the *D. gigas* genome provides insights into the integrated network of energy conserving complexes and structures present in this bacterium. Comparison with genomes of other *Desulfovibrio* spp. reveals the presence of two different CRISPR/Cas systems in *D. gigas*. Phylogenetic analysis using conserved protein sequences (encoded by *rpo*B and *gyr*B) indicates two main groups of *Desulfovibrio* spp, being *D. gigas* more closely related to *D. vulgaris* and *D. desulfuricans* strains. Gene duplications were found such as those encoding fumarate reductase, formate dehydrogenase, and superoxide dismutase. Complexes not yet described within *Desulfovibrio* genus were identified: Mnh complex, a v-type ATP-synthase as well as genes encoding the MinCDE system that could be responsible for the larger size of *D. gigas* when compared to other members of the genus. A low number of hydrogenases and the absence of the *codh/acs* and *pfl* genes, both present in *D. vulgaris* strains, indicate that intermediate cycling mechanisms may contribute substantially less to the energy gain in *D. gigas* compared to other *Desulfovibrio* spp. This might be compensated by the presence of other unique genomic arrangements of complexes such as the Rnf and the Hdr/Flox, or by the presence of NAD(P)H related complexes, like the Nuo, NfnAB or Mnh.

our research through the grant ADI/2006/M2.3/003 given to C.R.P. and O.F. We are also greatly indebted to STAB Vida and BIOCANT for their financial support. F.M.S (SFRH/BD/45211/2008), C.P. (SFRH/BPD/90823/2012) S.S. (grant SFRH/BPD/80244/2011), were supported by FCT fellowships. The work conducted in CPqRR – FIOCRUZ, was supported by Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES); Fundação de Amparo à Pesquisa do Estado de Minas Gerais, National Counsel of Technological and Scientific Development CNPq and Rede Integrada de Estudos Genômicos e Proteômicos (GENOPROT) through grants APQ-02382-10, PRI-00197-12, APQ-01085-12, and grants 476539/2010-2, 301652/2012-0 and 560943/2010-5.

Received: 20 March 2014; Revised: 30 April 2014; Accepted: 15 May 2014

***MicrobiologyOpen* 2014; 3(4): 513–530**

doi: 10.1002/mbo3.184

[a]These authors contributed equally to this study.

*This paper is dedicated to the memory of Professors Jean LeGall and António V. Xavier who have introduced us to the study of *Desulfovibrio gigas* and have greatly stimulated its research.

Introduction

Sulfate-reducing bacteria (SRB) are probably one of the most ancient forms of life on Earth. This group of anaerobic microorganisms, widespread in anoxic habitats, uses sulfate as main terminal electron acceptor to degrade organic compounds, with the consequent production of sulfide (Muyzer and Stams 2008). This process is extremely important in the sulfur and carbon cycles, since ~50% of the organic carbon mineralization in marine sediments is due to sulfate reduction (Jorgensen 1982). SRB are metabolically very versatile microorganisms, being able to use organic and inorganic substrates, as well as other short-chain fatty acids or ethanol for sulfate reduction. In recent years, new species were found to be able to grow on more diverse and less degradable substrates such as hydrocarbons or aromatic compounds (Rabus et al. 2006). Furthermore, due to the fact that many SRB use H_2 as an important substrate for sulfate reduction, they are able to participate in interspecies hydrogen transfer processes in synthropic communities with archaea (Walker et al. 2009; Plugge et al. 2010; Li et al. 2011). As a result of their metabolic flexibility, SRB can be found in almost all ecological environments on the planet. Moreover, these bacteria possess a wide biotechnological potential, especially in bioremediation of sulfate and heavy metals from natural environments and in removal of industrial waste liquids and sewage (Janssen et al. 2001; Lenz et al. 2008). On the other hand, due to the production of high amounts of hydrogen sulfide, SRB have large negative economic impact mainly as causative agents of microbial corrosion processes in anaerobic environments like those occurring in offshore oil production or waterlogged clay soils, resulting in economic losses (Hamilton 1985). Furthermore, they can create problems through a change in oil composition and souring of petroleum reservoirs (Huang and Larter 2005; Vance and Thrasher 2005).

Recent advances in genomics, biochemistry, and genetics of the SRB have greatly helped to identify the essential enzymes and complexes that participate in

sulfate respiration. The reduction of sulfate to sulfide during the respiratory process occurs in the cytoplasm. As such, electron transport chains and carriers must provide a link for the flow of the reducing equivalents ($[H^+]$ and electrons) between dehydrogenases and the terminal reductases (Rabus et al. 2006). Despite many efforts to understand the sites and mechanisms of energy conservation in sulfate respiration, the electron-transfer pathways that generate ATP from oxidative phosphorylation and create a proton gradient are not yet fully understood (Pereira et al. 2011). Most of the studies are focused on understanding the principles of sulfate reduction using *Desulfovibrio* genus. Among the various members of this genus, *Desulfovibrio gigas,* a curved rod bacterium, whose name was inspired by its unusual size (up to 11 μm) was for the first time isolated in 1963 by Jean LeGall from a water pond (LeGall 1963). After its isolation, this bacterium was used by many different groups to elucidate the structure of enzymes participating in energy transfer reactions such as hydrogenases, formate dehydrogenases, ferredoxins, cytochromes, and the xantine oxidase-related aldehyde oxido-reductase (molybdenum-containing aldehyde oxido-reductase; MOP) (Ambler et al. 1969; Romao et al. 1995; Volbeda et al. 1995; Matias et al. 1996; Frazao et al. 2000; Raaijmakers et al. 2002; Hsieh et al. 2005). Mechanistic and functional processes related to the energy metabolism and stress response have been also well studied in *D. gigas* (Silva et al. 2001; Broco et al. 2005; Rodrigues et al. 2006a; Morais-Silva et al. 2013). However, despite the accumulated information about this bacterium, a clear whole-genome context of the genes and metabolic complexes is not yet available for *D. gigas*. Previous analyses and comparison between the different species of SRB revealed that the composition of energy metabolism proteins, as well as stress-related proteins can vary quite significantly (Rabus et al. 2006; Pereira et al. 2008, 2011). *D. gigas* may, therefore, react to environmental cues and adapt to different environments by using different metabolic and structural components. Genome sequencing analysis is an important tool in order to fully understand which components may be involved in these adaptation and survival mechanisms. In this article, we examine the whole-genome sequence of this organism and perform a comparative genomic analysis with other *Desulfovibrionaceae*.

Materials and Methods

DNA sequencing, assembly, and annotation

DNA was isolated with the Wizard Genomic DNA Purification Kit (Promega, Mannheim, Germany). Sequencing was performed using a combination of several approaches: Sanger sequencing, using small fragment (2–6 kb) libraries; High throughput Roche Diagnostics 454 GS20 sequencing (Roche Diagnostics, Mannheim, Germany) (Keygene) and Illumina′s Solexa sequencing technology. Final gap closure was obtained either by primer walking or resequencing in the Personal Genome Machine (PGM) platform set up in STAB VIDA. The global coverage was 159.68-fold sequences. Ab initio assembly was performed using Velvet version 0.7.55 software (Zerbino and Birney 2008), and the consensus genomic sequence was obtained with Phrap (http://www.phrap.org/phredphrapconsed.html).

Structural annotation was performed using FgenesB (www.softberry.com), RNAmmer (Lagesen et al. 2007), tRNA-scan-SE (Lowe and Eddy 1997) and Tandem Repeat Finder (tandem.bu.edu/trf/trf.html). Functional annotation was performed by similarity, using public databases and InterProScan analysis (Zdobnov and Apweiler 2001). Protein-coding sequences were manually curated using Artemis (Rutherford et al. 2000). Comparative analyses for *Desufovibrio* spp. were performed using the BLAST–NCBI (Altschul et al. 1990) and InterProScan databases. The genomic and plasmidic sequences of *D. gigas* ATCC19364 were submitted to GenBank under the Acession No. CP006585 and CP006586, respectively.

Phylogenetic analysis

Evolutionary relationship between *Desulfovibrio* species was constructed using RpoB and GyrB concatenated sequences downloaded from GenBank (ftp://ftp.ncbi.nlm.nih.gov/). Sequence alignment was done using MAFFT software (Katoh et al. 2002) and LG evolutionary model (Le and Gascuel 2008) was selected for analysis using the ProtTest version 2. (Abascal et al. 2005). PhyML version 3.0 algorithm and the Maximum Likelihood method (Guindon et al. 2010) were used to create the phylogenetic tree. The evolutionary history of Cas1 proteins was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al. 1992).

In order to assess the number of genes shared between *D. gigas* and other *Desulfovibrio* species, a Venn diagram was built using the EDGAR database (https://edgar.computational.bio.uni-giessen.de/).

Codon usage analysis

The *D. gigas* codon usage was determined by using an EMBOSS tool called *cusp* (Rice et al. 2000). This program calculates a codon usage table for one or more nucleotide coding sequences (Table S2). The codon usage table and *D. gigas* coding sequences were next used in another EMBOSS program called *cai*, which calculates the Codon Adaptation Index (CAI) (Sharp and Li 1987).

Protein interaction network

A list of *D. gigas* proteins related to chemotaxis (Table S6) and oxygen response (Table S7) were used as query in the search

Table 1. General genome features of *Desulfovibrio gigas.*

Features	Value	% of total
Genome		
Genome size (bp)	3,693,899	100
DNA coding region (bp)	3,249,714	87.98
G + C content (bp)	2,341,530	63.39
Extracromossomal elements	1	
Number of replicons	1	
Total number of genes	3370	100
Stable rRNAS		
rRNAs	3	0.09
tRNAs	48	1.42
Protein-coding genes	3273	97.09
Genes density (bp/gene)	1128	
Average length of a gene (bp)	993	
Pseudogenes	47	1.39
Genes with assigned COG	2273	67.45
Selenocysteine-containing proteins	9	
Genes without assigned function	999	29.64
Poorly chracterized genes	395	11.72
Other elements		
CRISPR repeats	6	
Cas operons	2	
Transposases	17	
Mobile elements	1	

against the EDGAR database. The number of *D. gigas* orthologs found in 13 *Desulfovibrio* strains is depicted in radar graphs.

Results and Discussion

General genome features

The genome of *D. gigas* (CP006585) consists of one circular chromosome of 3,693,899 base-pairs (bp) having 3370 genes of which 3273 are protein-coding (see Table 1 and Fig 1A), classified according to its predicted COG function (Table S1). The genome has a G+C content of 63.4% that reflects a biased codon usage. Indeed, *D. gigas* prefers high G+C codons (66.87%), with a clear preference for cytosine (C) in the 3rd position (82.03% and Table S2). Indeed, among synonymous codons used by the 20 aminoacids, 13 of them are using more frequently one triplet ending by C. There are however, a few exceptions as leucine and valine which use respectively the CTG and GTG. The CAI calculated for all coding sequences has an average of 0.663, with a maximum of 0.863 (DGI_1104, a putative hydrolase) and the minimum of 0.198 (DGI_2086 and 3377, both hypothetical proteins). Genes encoding the hydrogenases (Table S24), as well as energy conserving transmembrane complexes (Table S25) present a higher CAI value, around 0.701, than the average, suggesting that these genes have a higher expression potential which was experimentally shown in the case of both the hydrogenases Ech and Hyn, (Morais-Silva et al. 2013).The genome is very compact as observed by its gene density of 1128 bp per gene and the average length of each gene is 993 bp. It contains 17 transposases (Table S3), whereas in other SRB genomes this number is in average 34 (Bennett 2004). This relative low number of transposable elements in *D. gigas* may indicate a low rate of reorganization of its genome. Other features include 47 pseudogenes and 48 tRNAs (Table 1), as well as 9 selenocysteine containing proteins (Table S4). Surprisingly, one single operon of rRNA was found in *D. gigas* in contrast to what was detected in other *Desulfovibrio* spp. that contain between 3 and 6 operons. The recently sequenced genome of the new strain *Salinarchaeum* sp. HArcht-Bsk1T, also contains one single rRNA operon (Dominova et al. 2013) as well as the bacterium *Mycobacteria,* a fact that was associated to their slow growth (Arnvig et al. 2005). The high generation time of *D. gigas* of around 8 h may be as well related to this fact. Besides, having in the genome solely 17 genes encoding transposases and only a single rRNA operon, may also indicate a decreased genome rearrangement, as multiple rRNA operons serves as sites for homologous recombination (Helm et al. 2003).

The plasmid of this bacterium (CP006586) has a size of 101,949 bp, containing 75 ORFs, of which 72 are coding regions (Table 2 and Fig 1B). Approximately one-third of the encoded polypeptides are annotated as hypothetical. Regarding the remaining annotated ORFs, the most representative functional group is composed of 12 proteins encoding acetyl, methyl, and glycosyl transferases. Interestingly, we could also identify an operon of 12 ORFs (DGIp_00010-00120) encoding a type II secretory system (T2SSs) which is involved in the secretion of folded and/or oligomeric exoproteins (Douzi et al. 2012). We have also identified a 30 kb operon encoding a set of capsule polysaccharide biosynthesis (*kps*) and transporter (*tag*) proteins. These features may indicate a mechanism used by *D. gigas* to secret and transport folded exoproteins. Another remarkable feature of *D. gigas* plasmid is related to the presence of the *apsK* gene encoding a bi-functional protein, predicted to have a sulfate adenylyltransferase and adenylylsulfate kinase activities (Marchler-Bauer et al. 2013).

Desulfovibrio gigas and its size

The size of *Desulfovibrio gigas* is larger than the one of other *Desulfovibrio* spp. Its length is of 5–10 μm and the width of 1.2–1.5 μm, whereas the other species have a cell size of 3–5 μm by 0.5–1 μm (Postgate and Campbell 1966). The bacterial morphogenesis and cell size are determined by the two major types of proteins, FtsZ, the

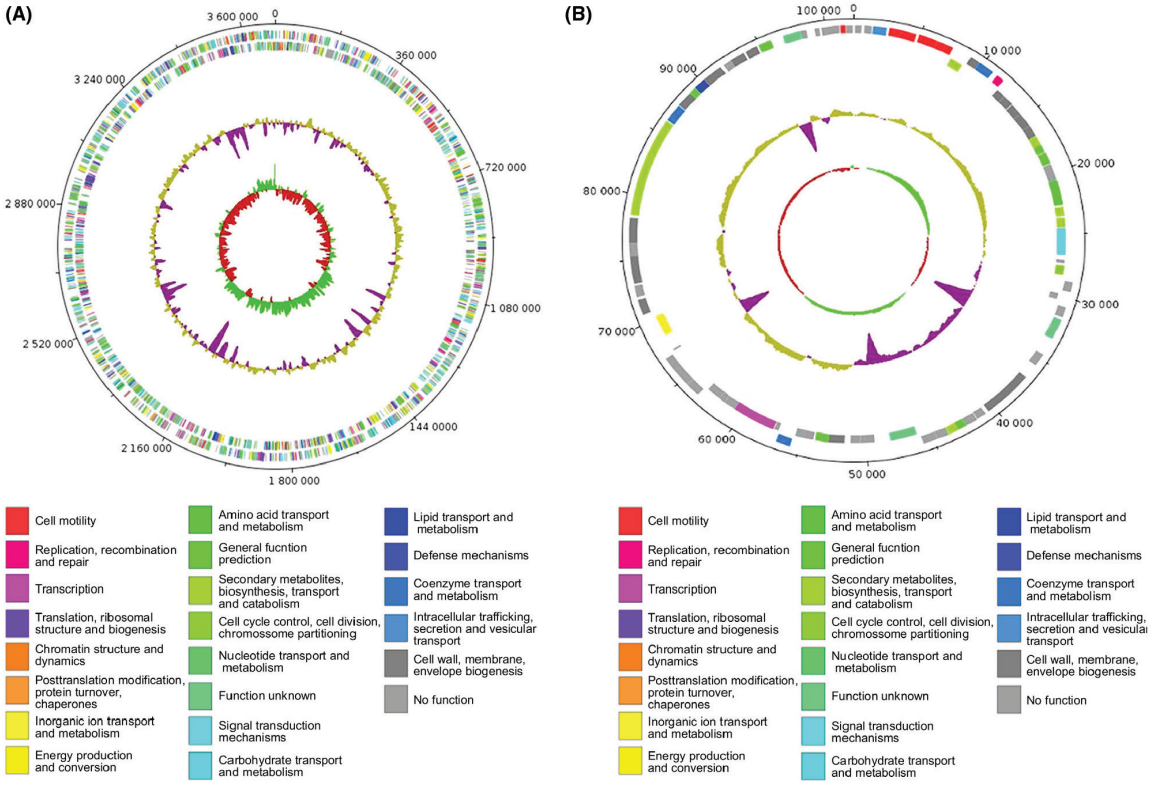


Figure 1. Structural representation of the circular chromosome (A) and plasmid (B) of *Desulfovibrio gigas*. Circular representations, from inside to the outside represent: (i) GC skew, richness of guanine over cytosine in the positive strand represented in green and cytosine over guanine represented in red; (ii) GC content, below average in purple, above average in gold; (iii) positive strand coding regions (below) and negative strand coding regions (above) colored according to COG functional terms of the best hit obtained from *Blastp* program; (iv) nucleotide position indicated in circular scale.

tubulin homolog responsible for cell division, and MreB, related to actin, which is involved in cell elongation of rod-shaped bacteria (Marshall et al. 2012).

D. gigas genome contains the inhibitor of the FtsZ assembly, the *minCDE* system similar to the one described for *E. coli* (DGI_3156, 3157 and 3158) (Fig. S1A) (de Boer et al. 1989), which is not detected in any other *Desulfovibrio spp* genomes so far sequenced. The Min system was described as participating in the accurate placement of the division site, allowing septum formation in the middle of the cell by inhibiting FtsZ polymerization. In fact, it was shown that the defects in the Min system components lead to a high frequency of aberrant FtsZ assembly at sites immediately adjacent to the cells poles (Rothfield et al. 2005; Marshall et al. 2012).

Table 2. General plasmid features of *Desulfovibrio gigas*.

Features	Value
Size (bp)	101,949
G + C contente (bp)	64,081
DNA coding region (bp)	79,425
Pseudogenes	3
Protein-coding genes	72
Gene density (bp/gene)	1415
Average length of a gene (bp)	1103

As *D. gigas* contains the *minCDE* genes, in contrast to other *Desulfovibrio* spp, this may suggest the involvement of the encoded polypeptides in the different size of this bacterium. Indeed, the presence of these genes may originate an inhibition of FtsZ assembly, leading to an increase in cell size. In addition to the Min system, a homolog of the nucleoid occlusion SmlA protein (DGI_2692), that prevents the polymerization of FtsZ and thus cell division, was also found (Bernhardt and de Boer 2005). We further detected a homolog of a third FtsZ assembly inhibitor that was described for *B. subtilis*, the *pgcA* gene (DGI_0235), which couples cell division to cell mass (Weart et al. 2007).

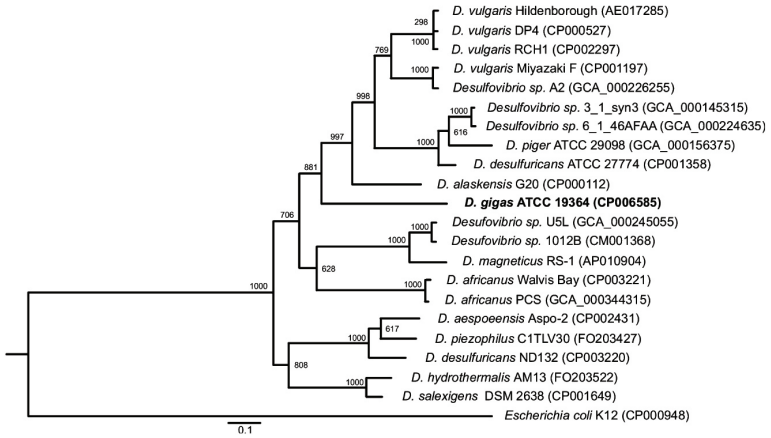


Figure 2. Evolutionary relationship of *Desulfovibrio* species. This tree was built based on RpoB and GyrB protein sequences using a Maximum Likelihood approach with 1000 iterations for the Bootstrap test, both implemented in the PhyML tool. The number at each node corresponds to the frequency of that branching occurred during the 1000 iterations. The sequences of the *E. coli* proteins were applied as outgroup. Accession numbers are indicated after the species names.

Regarding the MreB, considered as an organizer of cell wall synthesis, three genes encoding similar proteins appear in the *D. gigas* genome (DGI_0336, 0660 and 2254), whereas other *Desulfovibrio* spp. contain only two genes. Although the pathway by which MreB controls the cell width is not yet established, the presence of an extra *mreB* gene could as well contribute to the big size of *D. gigas*. As such, a putative interaction network of *D. gigas* proteins involved in the cell size, built based on the data obtained from *D. vulgaris* protein interactions using the STRING database (http://string-db.org/) and the data available in the literature (Bi and Lutkenhaus 1990; Weart et al. 2007; Fischer-Friedrich et al. 2010; Chien et al. 2012; Hill et al. 2012), can be drawn (Fig. S2).

Phylogenetic analysis of Desulfovibrio genus

A phylogenetic tree was built based on protein sequences coded by the conserved RpoB and GyrB protein sequences from 21 isolates of *Desulfovibrio* genus whose genomic sequences are available and annotated.

The analysis revealed two well-supported deep-branching main clades (Fig. 2). Within the upper clade, two groups emerge: one group contains *D. gigas* clustering with *D. alaskensis* G20, *D. piger* ATCC29098, *D. desulfuricans* ATCC27774, and *D. vulgaris* spp; the other group embraces *D. magneticus RS-1*, two *D. africanus* strains, and two not yet assigned *Desulfovibrio* species (Fig. 2). The lower clade contains a single group of *Desulfovibrio* species with many of them found in larger depths (piezophilic environment). The tree topology suggests a more divergent evolutionary history of the species included in the lower clade. In fact, gene structures associated with oxygen resistance and detoxification, such as the superoxide dismutase (SOD) genes (DGI_1536 and DGI_3082, Table S7), are present not only in *D. gigas* and in the closely related *D. vulgaris* species, but also in the subgroup containing *D. magneticus* RS-1. However, species observed in the lower clade, such as *D. piezophilus* and *D. hydrothermalis*, do not contain any homologous sequences for SOD genes. This different oxygen resistance gene structures could be the reflex of a different evolutionary process of this later group of *Desulfovibrio* spp since these species are found in environments where O_2 is present at very low levels (Ji et al. 2013).

Remarkably, according to this phylogenetic analysis, the isolates within *Desulfovibrio* genus not yet classified, namely *Desulfovibrio* sp. 3_1_syn3 together with *Desulfovibrio* sp. 6_1_46AFAA, *Desulfovibrio* sp. U5L along with *Desulfovibrio* sp. 1012B and *Desulfovibrio* sp. A2, are clustered with *D. desulfuricans, D. magneticus, and D. vulgaris*, respectively. Corroborating our data with respect to the *Desulfovibrio* sp. A2, using 16S rRNA gene sequence, a 99.1% overall sequence similarity with *D. vulgaris* Miyazaki was shown (Mancini et al. 2011). These findings indicate that they are closely related species and merit further investigation, in order to clarify their classification within the *Desulfovibrio* genus.

Another interesting aspect of this analysis relies in the positioning of *D. desulfuricans* ND132 within the lower clade of the phylogenetic tree, rather than in the upper clade, where *D. desulfuricans* appears (Fig. 2). This finding has already been observed by others and strongly

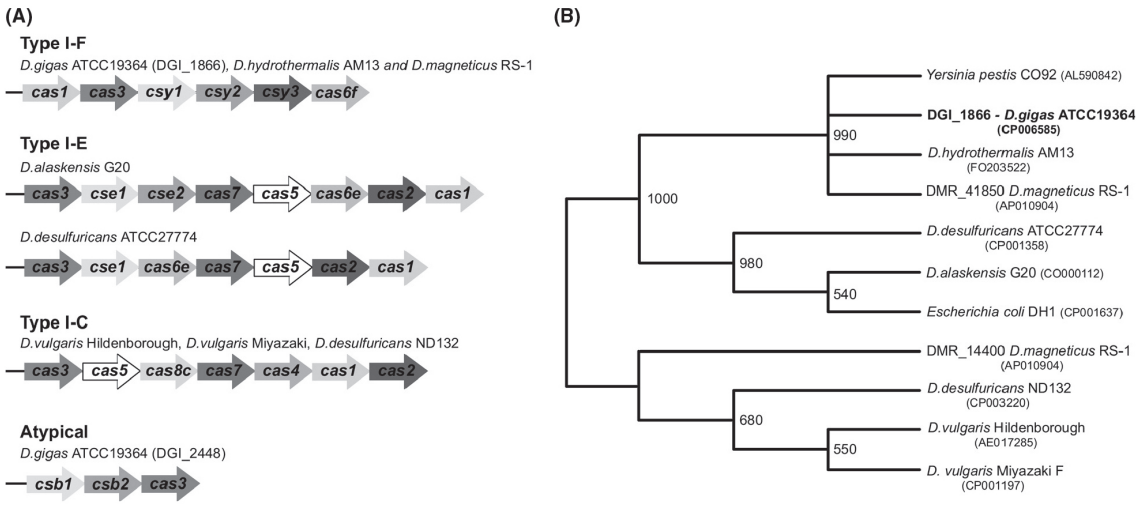


Figure 3. Distribution of different types of CRISPR/Cas systems among *Desulfovibrio* spp. (A) Operon structure of *cas* genes from the indicated *Desulfovibrio* spp. The operon organization was assessed using the DOE Joint Genome Institute (JGI) website (http://www.jgi.doe.gov/). Classification into the distinct Type I subtypes is according to (Makarova et al. 2011). (B) The evolutionary history of Cas1 proteins was inferred by using the Maximum Likelihood method. The bootstrap consensus tree inferred from 1000 replicates was taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches. Accession numbers are indicated after species name.

indicates that its classification should be reconsidered (Brown et al. 2011; Gilmour et al. 2011).

CRISPR/Cas systems in the *D. gigas* genome

CRISPRs are loci encompassing several short repeats functioning as an adaptive microbial immune system, that have also been shown to limit horizontal gene transfer (HGT) by preventing conjugation and plasmid transformation (Marraffini and Sontheimer 2008). Several types of CRISPR-associated proteins (Cas) are encoded by *cas* genes located in the vicinity of CRISPRs. Cas proteins are required for the multistep defense against intruder genetic elements. Their number, identity, and the corresponding operon organization appear to be extremely variable. Makarova et al. (2011) have proposed a classification of CRISPR/Cas systems in which the *cas1* and *cas2* genes constitute the core of three distinct types of system. Each system was further divided into different subtypes, on the basis of the gene composition and architecture of the respective operons.

In the particular case of *Desulfovibrio* spp., little is known about the presence of CRISPR sequences and Cas-associated genes. *D. vulgaris* Hildenborough appears to have a plasmidic CRISPR/Cas locus that falls into the subtype I-C system, according to the above mentioned classification criteria (see Fig. 3A and Makarova et al. (2011), Haft et al. (2005)). A survey of the genome of *D. gigas* for CRISPR repeats, revealed the presence of 6 CRISPR repeats with two of them being flanked by Cas operons (Table S5 and Fig 3A). One of the *D. gigas* CRISPR/Cas systems fall into the I-F type, for the first time reported in *Yersinia pestis*, and the other one does not fit in any of the known types of CRISPR/Cas systems (Fig. 3A).

Using a dedicated database (http://crispi.genouest.org/) (Rousseau et al. 2009), we searched for CRISPR sequences that have adjacent *cas* genes among the different species of *Desulfovibrio* genus. We have focused on CRISPR/Cas arrays that possess the ubiquitous core protein Cas1, which is involved in new spacer acquisition. We then used the conserved Cas1 protein as a scaffold to investigate the evolution of the CRISPR/Cas system in the *Desulfovibrio* genus (Fig. 3B). Remarkably, CRISP/Cas systems are absent from the genome of *D. aespoeensis* Aspo-2, *D. africanus* Walvis Bay, *D. piezophilus* C1TLV30, and *D. salexigens* DSM2638.

The phylogenetic tree of *Desulfovibrio* genus was used in order to explore the evolutionary bases of the CRISPR/cas loci (Fig 2). In the particular case of group I, the topology of Cas1 phylogenetic tree (Fig. 3B) together with the RpoB and GyrB based phylogeny of the genus *Desulfovibrio* (Fig 2), strongly suggests the divergence after speciation of an ancestor gene common to *D. gigas*

ATCC19364, *D. hydrotermalis* AM13, and *D. magneticus* RS-1. Furthermore, the Cas1 phylogeny shows *D. desulfuricans* ATCC27774 and *D. alaskensis* G20 grouping separately from the other *Desulfovibrio* spp. and of *E. coli* DH1 (Fig. 3B). These phylogenetic relationships together with the RpoB_GyrB phylogenetic tree indicate that CRISPR/Cas system I-E (group II) might have been acquired from HGT during prokaryotic evolution. Indeed, a comprehensive phylogenetic analysis of CRISPR/cas loci points toward their propagation via HGT events (Godde and Bickerton 2006). Regarding group III, it seems that the CRISPR/Cas subtype I-C is scattered across several *Desulfovibrio* spp. (Figs. 2, 3B). The absence of additional *Desulfovibrio* orthologues suggests that the acquisition of this CRISPR/Cas subtype may rely as well in HGT occurrences throughout evolution. Notably, *D. vulgaris* Hildenborough contains the CRISPR/cas locus in its megaplasmid, whereas the closely related *D. vulgaris* Miyazaki (Fig. 2) possesses a similar CRISPR/cas array in the chromosome. Godde and Bickerton have proposed that most megaplasmids should not be stably maintained in their host cells (Godde and Bickerton 2006). Consistently, the lack of a megaplasmid in *D. vulgaris* Miyazaki indicates that a recent HGT event might have been responsible for the appearance of CRISPR locus in *D. vulgaris* Hildenborough.

Strategies to survive oxygen and nitric oxide

SRB, in the diverse environmental niches they occupy, can come across with reactive oxygen or nitrogen species that cause oxidative damage to the cells. Formerly classified as strict anaerobes there is, however, growing evidence that they are able to cope with oxygen and to use it to produce ATP even if they are unable to grow in its presence. As such, the organisms have developed several strategies to avoid such damage.

The response to different oxygen concentrations in microorganisms, aerotaxis, is often initiated by the transmembrane chemoreceptors, the methyl-accepting chemotaxis proteins, and involves many other proteins organized in a cascade of reactions activating the flagellar motor, allowing the cells to move to an optimal oxygen gradient (Armitage 1997). SRB within the microbial mats and oxic environments are motile, and active movements are observed in response to change in oxygen gradients which were interpreted as a strategy to survive in these environments (Krekeler et al. 1989; Canfield and Des Marais 1991; Teske et al. 1998; Eschemann et al. 1999). The sensing of extra and/or intracellular signals is followed by their transduction to the transcriptional and post-transcriptional machineries. As it was previously demonstrated, *D. gigas* contains an operon encoding the chemotaxis proteins CheB, CheR, CheW, CheY, and CheA, that are co-transcribed as an 11 kb mRNA whose expression is not altered either by O_2 or nitric oxide (NO, Felix et al. 2006). By searching *D. gigas* genome, many other chemotaxis coding regions were found scattered throughout the genome (Table S6). A comparison of the newly identified genes coding for chemotaxis proteins against other sequenced *Desulfovibrio* spp., indicate that few of these operons have orthologous in closely related species such as *D. vulgaris* or *D. desulfuricans* strains (Fig. 4A). As such, it is clear that the genes without orthologs represent specific mechanisms that *D. gigas* uses to sense and avoid unfavorable aerobic conditions. Strikingly, the closely related *D. vulgaris* DP4 and RCH1 as well as *D. desulfuricans* are those among the *Desulfovibrio* spp. that have fewer orthologs genes encoding chemo-

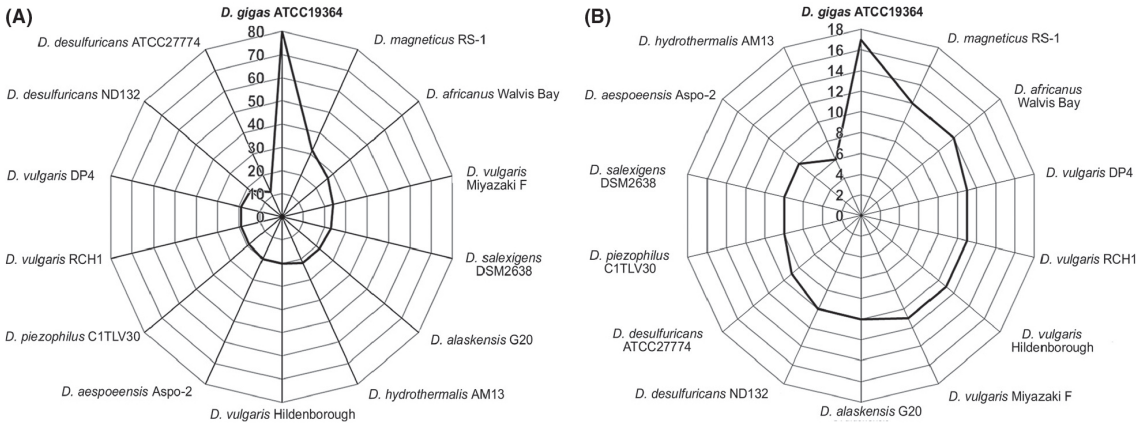


Figure 4. Radar Graphs comparing orthologs of *D. gigas* within the genome of sequenced *Desulfovibrio* spp. (A) Genes involved in chemotaxis response; (B) genes involved in O_2 sensing; Orthologs search was conducted using the EDGAR database.

taxis polypeptides when compared to *D. gigas* (See Fig. 4A).

Besides sensing, these microorganisms have developed a network of defense mechanisms against reactive oxygen species (ROS), being the toxic O_2 eliminated by dismutation to H_2O_2 and O_2, a reaction catalyzed by the SOD (dos Santos et al. 2000). The accumulation of toxic H_2O_2 is further eliminated by the catalase which is found in *D. gigas* genome as a single gene (DGI_2858) (dos Santos et al. 2000). *D. gigas* contains in its genome two SOD genes, one named neelaredoxin and another one (DGI_1536) here described for the first time (see Table S7). Neelaredoxin from *D. gigas* was shown to be a bifunctional protein that has both superoxide reductase and SOD activities. (Silva et al. 1999; Abreu et al. 2002). *D. gigas* genome also contains genes encoding three rubrerythrins, one peroxiredoxin, one rubredoxin-like protein, and three F390 synthetase proteins (Table S7), which have been shown to be related to defense mechanisms against oxidative stress.

As illustrated in the radar chart, 17 genes are involved in O_2 metabolism of *D. gigas* some of which have orthologs in other species of *Desulfovibrio*. As such, *D. gigas* shares 12 genes with *D. magneticus* RS-1 and 6 genes with *D. hydrothermalis* AM13, a more distant species (Fig. 4B). Interestingly, when observed in more detail, the species grouped together with *D. hydrothermalis* AM-13 in the phylogenetic analysis (Fig. 2), such as *D. aespoeensis* and *D. salexigens* showed an increased number of superoxide reductases (two genes) when compared to *D. gigas* or *D. vulgaris*, that only possesses one gene, according to the SORGOdb database (Lucchetti-Miganeh et al. 2011).

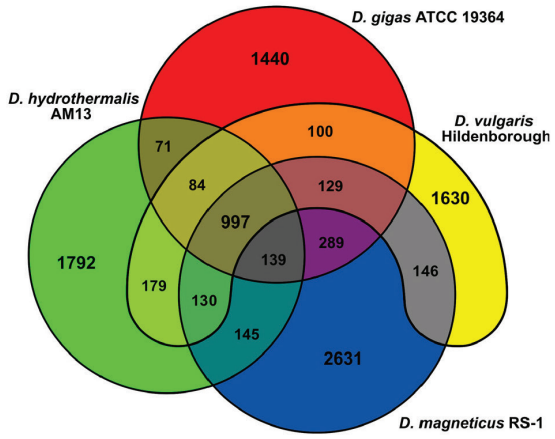


Figure 5. Protein orthology comparison among genomes of *Desulfovibrio gigas, D. magneticus* RS-1, *D. hydrothermalis* AM 13, and *D. vulgaris* Hildenborough. The Venn diagram shows shared ortholog groups for any given gene of each species under analysis.

Furthermore, it is also interesting to notice that *D. gigas* presents a higher number of orthologous proteins regarding O_2 sensing and detoxification with *D. magneticus* RS-1 and *D. africanus* Walvis Bays species than with the more closely related *D. alaskensis* G20 and *D. vulgaris* (Fig. 4A and B). A whole-genome orthology analysis using the EDGAR database confirms this fact, as in general, a higher number of orthologous groups are observed between *D. gigas* and *D. magneticus* RS-1 than with *D. vulgaris* Hildenborough (Fig 5). This result was not expected on the basis of the phylogenetic results obtained, since *D.gigas* is more closely related to *D.vulgaris* Hildenborough than to *D.magneticus* RS-1. It is possible that only specific groups of genes included in the category of chemotaxis and detoxification show similarity to *D. magneticus* RS-1.

Another mechanism of O_2 detoxification involves the participation of the flavodiiron protein, rubredoxin:oxygen reductase (ROO) (Chen et al. 1993), which was also shown to protect *D. gigas* against nitrosative stress by acting as a NO reductase (Rodrigues et al. 2006b). Under nitrosative stress, *roo* transcription is regulated by NorR (NorR1L). A second putative *norR* gene designated as *norR2L* was found in *D. gigas* genome (Table S8) although its function is still unknown (Varela-Raposo et al. 2013). *D. gigas* genome also includes one copy of 'hybrid cluster protein' (HCP), a protein with an unusual structure (Cooper et al. 2000) proposed to have a function in nitrogen cycle due to its hydroxylamine reductase activity (Wolfe et al. 2002; Cabello et al. 2004; Overeijnder et al. 2009). A role in defense against oxidative stress has also been suggested for HCP on the basis of its peroxidase activity (Almeida et al. 2006). While in other *Desulfovibrio* spp., HCP is co-expressed with a hypothetical ferredoxin (*frdx*) gene (Rodionov et al. 2004) in *D. gigas* it is encoded by a monocistronic gene (Fig. S1B). It is also important to mention that *hcpR*, a gene encoding a transcriptional regulator of *hcp* expression identified in other *Desulfovibrio* spp., was also observed in *D. gigas* upstream of *hcp* although localized in opposite direction (Table S9, Fig. S1B) (Cadby et al. 2011). *D. gigas* genome encodes also the membrane complex cytochrome *c* nitrite reductase (NrfHA), which is suggested to play a role in nitrite detoxification since no growth on nitrite or nitrate is reported for *D. gigas*, as well as for *D. vulgaris*. (Greene et al. 2003; He et al. 2006). Other nitrate reductases as well as nitroreductases encoded in *D. gigas* genome (Table S8) might be involved in NO detoxification mechanisms.

Central carbon metabolism

D. gigas accumulates large amounts of polyglucose as an endogenous carbon and energy reserve, utilizing these

sugar compounds for growth (Fareleira et al. 1997). We have conducted a broad analysis in its genome to identify the elements of the central carbon metabolism involved in many different pathways (Table S11 to S17). Biochemical studies have shown (Fareleira et al. 1997), that *D. gigas* contains all the genes encoding proteins of the Embden-Meyerhof pathway (Table S13), whereas the genes coding for the hexokinase and the 2-keto-3-deoxygluconate 6-phosphate (KDGP) aldolase of the Entner-Doudoroff pathway are lacking (Table S14). *D. gigas* belongs to SRB group of incomplete-oxidizers, producing acetate and CO_2 as its main end-products from substrate oxidation. Inspection of the genome reveals that the genes corresponding to 2-oxoglutarate dehydrogenase, 2-oxoglutarate synthases, and both subunits of the succinyl Co-A ligase, *sucC* and *sucD* (Table S15) from the tricarboxylic acid (TCA) cycle are absent. Both copies of the succinate:quinone oxidoreductase (SQR), one of each is identified here (DGI_0826 to DGI_0828 - Table S16), appear to function mainly as fumarate reductases rather than as succinate dehydrogenases, due to a conserved glutamine residue (Glu180) in the Sdh/FdrC subunit (Zaunmuller et al. 2006).These results indicate that both oxidative and reductive TCA cycle pathways are not fully functional and are likely to have a biosynthetic function, as suggested for *D.vulgaris* Hildenborough (Heidelberg et al. 2004). In the Wood-Ljungdahl pathway (Table S17) the genes coding for a key element from this pathway (Ragsdale and Pierce 2008), the bifunctional carbon monoxide dehydrogenase/acetyl-CoA synthase (CODH/ACS) enzyme, are absent in *D. gigas*. Instead, like some *Desulfovibrio* spp. such as *D. magneticus* and *D. africanus* strains, *D. gigas* genome codes for an aerobic-type CODH of the *coxSLM* type, similar to the CO dehydrogenase of *Oligotropha carboxidovorans* (Dobbek et al. 1999). This enzyme shows a high sequence similarity with the aldehyde oxidoreductase (MOP) from *D. gigas* itself (Romao et al. 1995). This may suggest that this CO dehydrogenase could play a function in oxygen metabolism and resistance in *D. gigas* rather than being part of the Wood-Ljungdahl pathway as is the case of *D. vulgaris* Hildenborough, which presents a *codh/acs* gene. Furthermore, the absence of the bifunctional enzyme in *D. gigas* indicates that in contrast to *D. vulgaris* Hildenborough, CO cycling (Voordouw 2002) is not an effective mechanism of energy conservation.

Energy metabolism

A survey of *D. gigas* genome revealed several genes encoding dehydrogenases that oxidize organic acids and alcohols, as well as putative transporters and permeases for these substrates (Tables S18 to S20). Pyruvate, the main metabolic intermediate of organic carbon oxidation can be oxidized by the two pyruvate oxidoreductases (DGI_0996 and DGI_1712/DGI_1713) as well as by other oxo-organic acid ferredoxin: oxidoreductases enzymes present (Table S21). Although *D. gigas* genome reveals many genes encoding such complexes, the pyruvate:formate lyase (*pfl*), a gene involved in fermentative metabolism, was not identified. This enzyme produces acetyl-CoA and formate when pyruvate is the main carbon and energy source. As suggested for *D. vulgaris* Hildenborough, formate cycling could contribute to energy conservation in a mechanism similar to CO or hydrogen cycling (Voordouw 2002; Heidelberg et al. 2004). The apparent absence of this gene in *D. gigas* suggests that formate cycling is not occurring although this bacterium is able to grow using formate as the main electron donor (our unpublished results), since it presents two genes encoding formate dehydrogenases (Table S20). One of these enzymes, a tungsten seleno-protein, was already described (Almendra et al. 1999), whereas the second has not been reported to our knowledge (DGI_3334 and DGI_3335).

As other *Desulfovibrio* spp., *D. gigas* grows chemolithotrophically deriving energy from hydrogen oxidized in the periplasm by hydrogenases, coupled to sulfate reduction in the cytoplasm, creating a proton gradient ultimately used to generate ATP through F_1F_0-ATP synthase (Table S22). The electrons generated in the periplasm, by periplasmic hydrogenase activity, are transferred through the membrane for the sulfate reduction, in the cytoplasm, by multiheme c_3-type cytochromes (at the periplasmic side) and membrane-bound electron transport complexes.

The presence of at least three c_3-type cytochromes was found in *D. gigas* genome (Table S23). The full set of genes necessary for the dissimilatory sulfate reduction to sulfide were also detected, as well as specific sulfate permeases (Table S10). Interestingly enough, in the case of the ATP-synthase, not only the genes encoding the F_1F_0-ATP synthase were identified (Table S22) but another ATP-synthase, which apparently is not present in other *Desulfovibrio* spp, was found (Fig. S1A). This enzyme is similar to the Vacuolar-type ATPases (V_oV_1) and in some anaerobic bacteria, such as *Enterococcus hirae*, it functions as a sodium pump (Kakinuma et al. 1999). In *D. gigas,* this second ATPase could enhance ATP production derived from transmembrane electrochemical proton gradient.

In contrast to other *Desulfovibrio* spp. genomes so far sequenced (Pereira et al. 2011), only two [NiFe] type hydrogenase are present in *D. gigas*: the periplasmic HynAB (Volbeda et al. 1995) and the energy conserving Ech hydrogenase (Rodrigues et al. 2003) (Table S24). Recent work performed using mutant strains for these genes indicates that, although it is possible that the hydrogen cycling model of energy conservation (Odom

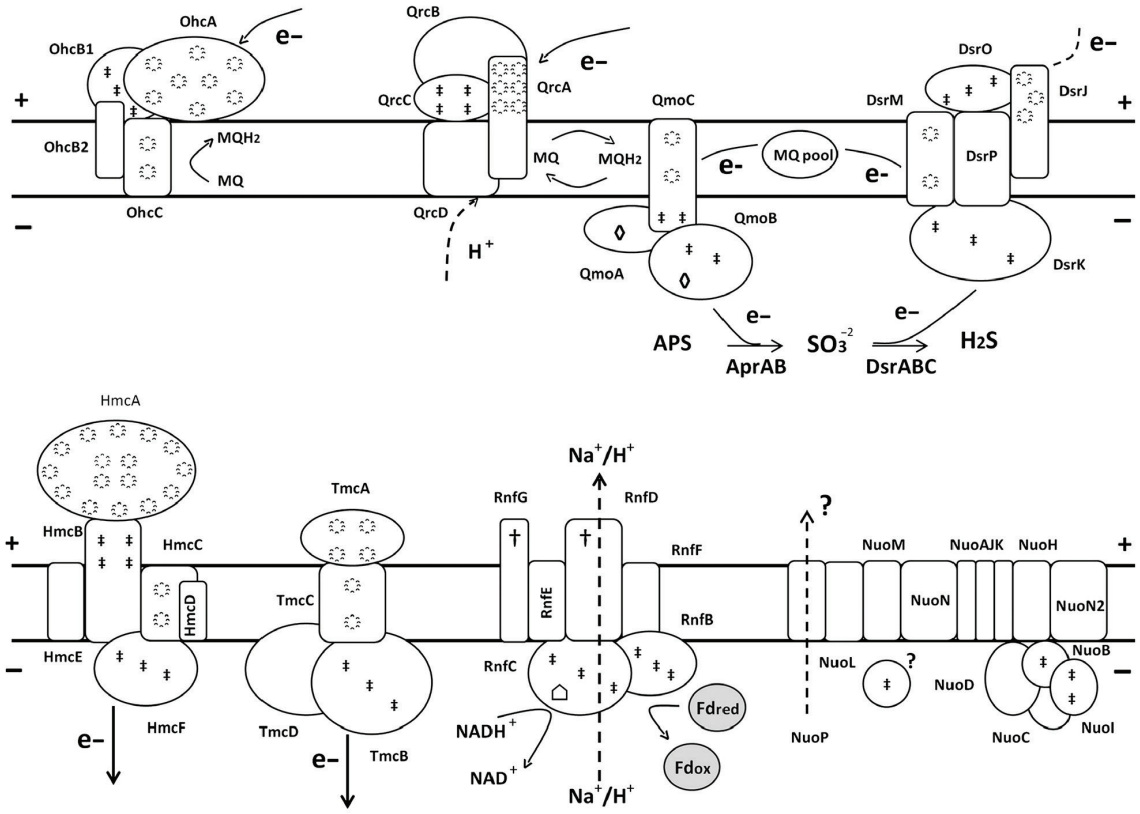


Figure 6. Schematic representation of membrane-bound electron-transfer complexes present in *D. gigas* genome. The complexes were identified in the genome according to their predicted function: quinone reduction, Ohc and Qrc; quinol oxidation, Qmo; transmembrane electron transfer/sulfite reduction DsrMKJOP, Hmc and Tmc; and NADH/Fd oxidation, Rnf and Nuo. Symbols represent:: ⁂, heme; ‡, iron sulfur center; †, FMN cofactor; ⌂, flavin cofactor. and ◊, FAD cofactor. Dashed lines represent hypothetical pathways for electron/proton flow.

et al. 1981) is effective, it appears to contribute substantially less to the final energy yield *of D. gigas* as proposed for other *Desulfovibrio* spp. (Morais-Silva et al. 2013). This could be a reflex of the unusual low number of these enzymes in *D. gigas*. An attempt to generate a double mutant strain in both hydrogenases (unpublished data) indicated that at least one of these hydrogenases might be essential for cell viability. Indeed, the double mutant was unable to grow in diverse respiratory and fermentative conditions.

Energy conservation

Sulfate reducers contain several transmembrane redox complexes involved in energy metabolism and conservation (Pereira et al. 2011) (Fig 6 and Table S25). The genome of *D. gigas* encodes two transmembrane multiheme cytochrome *c* complexes, Tmc and Hmc, described as participating in electron transfer from periplasmic hydrogen oxidation to sulfite reduction as transmembrane electron circuits (Rossi et al. 1993; Pereira et al. 2006). An octa-haem cytochrome c complex (Ohc), proposed to transfer electrons from the periplasm to the quinone pool, due to the absence of the cytoplasmic CCG protein, was also observed. Furthermore, we identified the quinone interacting membrane-bound oxidoreductase complex (*qmoABC*) and the transmembrane electron transfer DsrMKJOP complex, both related to sulfate reduction and suggested to act in the electron transfer to the final reductases, Apr (*aprAB*) and Dsr (*dsrABC*), respectively (Pires et al. 2003; Dahl et al. 2005). The presence of the Qrc (*qrcABCD*) quinone reduction complex, which was shown to transfer electrons from the Tpl-c_3 cytochrome to the menaquinone during sulfate respiration in a quinone:menaquinone loop together with the Qmo complex (Venceslau et al. 2010), suggests the existence in *D. gigas* of a mechanism of energy conservation linking periplasmic hydrogen or formate oxidation to cytoplasmic sulfate

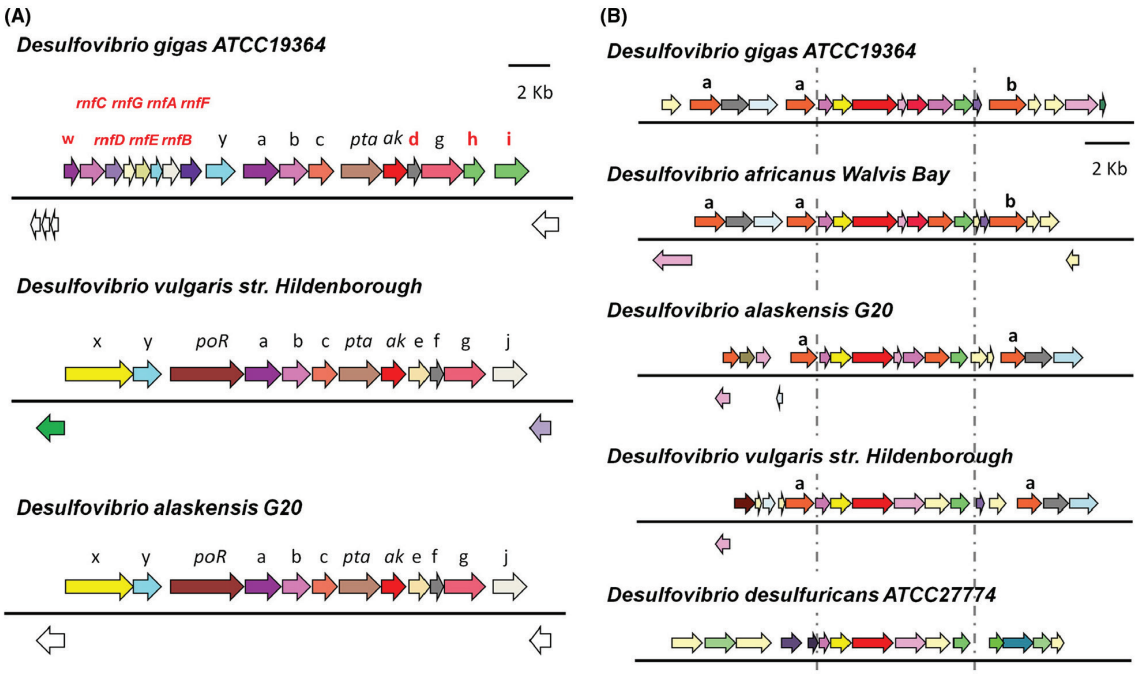


Figure 7. Genomic organization of the operons involved in the energy metabolism of *D.gigas*. (A) Comparison of the *Desulfovibrio spp* genomic regions containing the D-Lactate dehydrogenase operon. Differences in genomic organization are with red letters and genes are indicated as follows: *poR* - Pyruvate-ferredoxin oxidoreductas; *pta* - phosphate acetyltransferase; *ak* - acetate kinase; *rnfC, rnfD, rnfE, rnfG, rnfA, rnfB, rnfF* - Rnf complex; *w* - cytochrome c type protein; *x* - response regulator; *y* - sigma-54 response regulator; *a*, *b* and *c* - cytochrome c, lactate permease, and iron sulfur cluster protein (Ldh1a) subunits of the D-lactate dehydrogenase,; *d* - lactate utilization protein B/C; e and f - hypothetical proteins; *g* - iron sulfur cluster binding protein (Ldh1b); *h* - FMN-dependent α-hydroxyacid dehydrogenase; *i* - Sodium-dependent transporter; *j* - methyl-accepting chemotaxis protein. Identification of the single letter genes was made according to BLAST best hit value. Gene cluster organization of *D.vulgaris* Hildenborough and *D.alaskensis* G20 were obtained at the DOE Joint Genome Institute: http://www.jgi.doe.gov/. (B) Organization of genomic regions containing the HdrABC/FloxABCD operon and neighboring genes. Hdr/Flox operon appears between dashed lines and is composed of three subunits of Hdr-like proteins and four subunits of the flavin oxidoreductase genes (Flox) in *D.gigas*, *D.africanus,* and *D.alaskensis* or three subunits in *D.vulgaris* Hindelborough and *D.desulfuricans* ATCC27774. α-alcohol dehydrogenase; β-aldehyde dehydrogenase.

reduction. In addition, complexes involved in NAD(P)H and ferredoxin oxidation were identified (Table S25). An operon coding for the NADH:quinone oxidoreductase (*nuo*), firstly reported in *D. magneticus* RS-1 (Nakazawa et al. 2009) was also detected. This enzyme complex is proposed to couple NADH oxidation to proton translocation (Spring et al. 2012). However, the genes encoding the NADH dehydrogenase module (*nuoEFG*) are absent, suggesting a different electron donor, such as ferredoxin (Fd), instead of NADH (Pereira et al. 2011). Notably, a complex with high similarity to the *nuo* complex, the Mnh Na+/H+ antiporter, that was not detected in other *Desulfovibrio* spp. genomes, is present in *D. gigas* (Fig. S1A). This complex is suggested to function as a transmembrane electron potential-generating NADH dehydrogenase rather than as a secondary transmembrane electron potential-consuming antiporter, directly accounting for the great transmembrane electron potential in *Staphylococcus aureus* (Bayer et al. 2006). The presence of a similar mechanism in *D. gigas* might compensate for the apparent lack of energy conservation through metabolite cycling mechanisms, such as CO, formate or hydrogen cycling, deduced from its genome.

A search of the *D. gigas* genome also revealed the presence of the Rnf complex (*rnfCDGEABF*), proposed to function as a Na^+-translocating, ferredoxin:NAD^+ oxidoreductase (Biegel and Muller 2010) and a multiheme cytochrome *c* in the same operon (Fig 7A–w), hypothesized to mediate the electron transfer between the periplasmic cytochrome *c* pool and the cytoplasmic NAD(P)H/Fd (Li et al. 2006; Pereira et al. 2011). Another gene with similarity to cytochrome *c* is found adjacent to the Rnf complex in *D. gigas*, corresponding to cytochrome c subunit of D-lactate dehydrogenase (Fig. 7A– a). Interest-

ingly, the *rnf* operon is not present in the genomic context of this dehydrogenase in other *Desulfovibrio spp.*, being replaced by the pyruvate:oxidoreductase (*poR*). This fact may indicate that the Rnf complex in *D. gigas* could be directly involved in the electron transport from lactate to Fd/NADH or between these two elements.

Another group of energy-conserving enzymes and complexes are those related to electron bifurcation processes. *D. gigas* genome encodes two paralogous (Table S26) heterodimeric transhydrogenase (NfnAB), responsible for the reversible NADH-dependent reduction of $NADP^+$ by Fd (Wang et al. 2010).

Only one cytoplasmic hydrogenase was observed in *D. gigas* genome. We have, however observed a sequence of an electron bifurcating complex: the HdrABC/FloxABCD (Fig 7B). Flox gene products are likely to oxidize NAD(P)H and transfer electrons to the HdrABC proteins (Pereira et al. 2011) (Table S27). These genes are found in other *Desulfovibrio* spp., such as *D. vulgaris* and *D. alaskensis*, between two alcohol dehydrogenases (Fig 7B– a), suggesting that they might be involved in the electron transfer from alcohol substrates. The presence of an aldehyde dehydrogenase (Fig. 7B– b), found downstream of this operon in *D. gigas*, as well as *D. africanus* Walvis Bay, might indicate that this complex could also use aldehydes as another electron source to this complex. This genomic arrangement suggests that not only alcohol but also aldehyde oxidation could participate in mechanisms of energy conservation in *D. gigas*.

Conclusions

The observations reported for the genome of *D. gigas* ATCC19364 highlight the differences found within several species of the *Desulfovibrio* genus. The larger size of *D. gigas* cells when compared to other *Desulfovibrio* spp. might be a reflex of the presence of FtsZ inhibitors, such as the MinCDE system, which was not described for any members of this genus. In accordance, the presence of a single rRNA operon and multiple CRISPR/Cas elements specific for this species might be involved in the phylogenetic separation of *D. gigas*, placing it more closely related to *D. vulgaris* and *D. desulfuricans* strains. However, the presence of a different composition of genes involved in certain metabolic aspects, like sensing and response to oxygen and NO stress, highlighted by the presence of a new SOD, a second *norR* transcriptional factor (NorRL2), several putative nitrate reductases and an aerobic-type CODH, reveal a greater number of orthologous groups with more distant related species like *D. magneticus*. This also indicates a highly developed and flexible enzymatic machinery to overcome the deleterious effects of an aerobic environment. This flexibility can be further detected in the genes involved in the energy metabolism and conservation, as new proteins (Fdr and Fdh) and complexes, such as a secondary vacuolar-type ATPase and two complexes linking NAD(P)H and ferredoxins with electron transfer (Nuo and Mnh) were identified. On the other hand, a low number of hydrogenases and the absence of *codh/acs* and *pfl genes* indicate that the intermediate compounds (H_2, CO, and formate) do not contribute to mechanisms of energy conservation in *D. gigas* as much as they do in other *Desulfovibrio* spp. Despite that, recent experimental analysis performed using mutants for genes encoding hydrogenases demonstrates that at least one hydrogenase is required for cell viability. Interestingly, specific genomic elements, like the presence of a cytochrome *c* in the Rnf complex and an aldehyde dehydrogenases in the vicinity of the Hdr/Flox operon may provide alternative routes for energy conservation processes, that could compensate the absence of the above mentioned genes or multiple hydrogenases. This might indicate that different substrates (alcohols and aldehydes) and coenzymes ($NAD^+/NADP^+$) could play a more important role in redox reactions of *D. gigas* than previously thought.

Acknowledgments

We thank Teresa Barata and Mario Vicente for their support in the first part of this work. We also would like to thank Edson Luiz Folado, Leilane Oliveira Gonçalves, and Elvira C. A. Horácio for their help in processing the genomic data. This work was supported by Fundação para Ciência e Tecnologia (FCT) through grants PTDC/BIA-IC/104030/2008 given to C. R. P., Pest-OE/EQB/LA0004/2011 given to ITQB. Agência de Inovação (ADI) also supported our research through the grant ADI/2006/M2.3/003 given to C. R. P. and O. F. We are also greatly indebted to STAB Vida and BIOCANT for their financial support. F.M.S (SFRH/BD/45211/2008), C. P. (SFRH/BPD/90823/2012) S. S. (grant SFRH/BPD/80244/2011), were supported by FCT fellowships. The work conducted in CPqRR – FIOCRUZ, was supported by Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES); Fundação de Amparo à Pesquisa do Estado de Minas Gerais, National Counsel of Technological and Scientific Development (CNPq), and Rede Integrada de Estudos Genômicos e Proteômicos (GENOPROT) through grants APQ-02382-10, PRI-00197-12, APQ-01085-12, and grants 476539/2010-2, 301652/2012-0, and 560943/2010-5.

Conflict of Interest

None declared.

References

Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Abreu, I. A., A. V. Xavier, J. Le Gall, D. E. Cabelli, and M. Teixeira. 2002. Superoxide scavenging by neelaredoxin: dismutation and reduction activities in anaerobes. J. Biol. Inorg. Chem. 7:668–674.

Almeida, C. C., C. V. Romao, P. F. Lindley, M. Teixeira, and L. M. Saraiva. 2006. The role of the hybrid cluster protein in oxidative stress defense. J. Biol. Chem. 281:32445–32450.

Almendra, M. J., C. D. Brondino, O. Gavel, A. S. Pereira, P. Tavares, S. Bursakov, et al. 1999. Purification and characterization of a tungsten-containing formate dehydrogenase from *Desulfovibrio gigas*. Biochemistry 38:16366–16372.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Ambler, R. P., M. Bruschi, and J. le Gall. 1969. The structure of cytochrome c'(3) from *Desulfovibrio gigas* (NCIB 9332). FEBS Lett. 5:115–117.

Armitage, J. P. 1997. Behavioural responses of bacteria to light and oxygen. Arch. Microbiol. 168:249–261.

Arnvig, K. B., B. Gopal, K. G. Papavinasasundaram, R. A. Cox, and M. J. Colston. 2005. The mechanism of upstream activation in the rrnB operon of *Mycobacterium smegmatis* is different from the *Escherichia coli* paradigm. Microbiology 151:467–473.

Bayer, A. S., P. McNamara, M. R. Yeaman, N. Lucindo, T. Jones, A. L. Cheung, et al. 2006. Transposon disruption of the complex I NADH oxidoreductase gene (snoD) in *Staphylococcus aureus* is associated with reduced susceptibility to the microbicidal activity of thrombin-induced platelet microbicidal protein 1. J. Bacteriol. 188:211–222.

Bennett, P. M. 2004. Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement. Methods Mol. Biol. 266:71–113.

Bernhardt, T. G., and P. A. de Boer. 2005. SlmA, a nucleoid-associated, FtsZ binding protein required for blocking septal ring assembly over chromosomes in *E. coli*. Mol. Cell 18:555–564.

Bi, E., and J. Lutkenhaus. 1990. Interaction between the min locus and ftsZ. J. Bacteriol. 172:5610–5616.

Biegel, E., and V. Muller. 2010. Bacterial Na+-translocating ferredoxin:NAD+ oxidoreductase. Proc. Natl Acad. Sci. USA 107:18138–18142.

de Boer, P. A., R. E. Crossley, and L. I. Rothfield. 1989. A division inhibitor and a topological specificity factor coded for by the minicell locus determine proper placement of the division septum in *E. coli*. Cell 56:641–649.

Broco, M., M. Rousset, S. Oliveira, and C. Rodrigues-Pousada. 2005. Deletion of flavoredoxin gene in *Desulfovibrio gigas* reveals its participation in thiosulfate reduction. FEBS Lett. 579:4803–4807.

Brown, S. D., C. C. Gilmour, A. M. Kucken, J. D. Wall, D. A. Elias, C. C. Brandt, et al. 2011. Genome sequence of the mercury-methylating strain *Desulfovibrio desulfuricans* ND132. J. Bacteriol. 193:2078–2079.

Cabello, P., C. Pino, M. F. Olmo-Mira, F. Castillo, M. D. Roldan, and C. Moreno-Vivian. 2004. Hydroxylamine assimilation by *Rhodobacter capsulatus* E1F1. requirement of the hcp gene (hybrid cluster protein) located in the nitrate assimilation nas gene region for hydroxylamine reduction. J. Biol. Chem. 279:45485–45494.

Cadby, I. T., S. J. Busby, and J. A. Cole. 2011. An HcpR homologue from *Desulfovibrio* desulfuricans and its possible role in nitrate reduction and nitrosative stress. Biochem. Soc. Trans. 39:224–229.

Canfield, D. E., and D. J. Des Marais. 1991. Aerobic sulfate reduction in microbial mats. Science 251:1471–1473.

Chen, L., M. Y. Liu, J. Legall, P. Fareleira, H. Santos, and A. V. Xavier. 1993. Purification and characterization of an NADH-rubredoxin oxidoreductase involved in the utilization of oxygen by *Desulfovibrio gigas*. Eur. J. Biochem. 216:443–448.

Chien, A. C., N. S. Hill, and P. A. Levin. 2012. Cell size control in bacteria. Curr. Biol. 22:R340–R349.

Cooper, S. J., C. D. Garner, W. R. Hagen, P. F. Lindley, and S. Bailey. 2000. Hybrid-cluster protein (HCP) from *Desulfovibrio vulgaris* (Hildenborough) at 1.6 Å Resolution. Biochemistry 39:15044–15054.

Dahl, C., S. Engels, A. S. Pott-Sperling, A. Schulte, J. Sander, Y. Lubbe, et al. 2005. Novel genes of the dsr gene cluster and evidence for close interaction of Dsr proteins during sulfur oxidation in the phototrophic sulfur bacterium *Allochromatium vinosum*. J. Bacteriol. 187:1392–1404.

Dobbek, H., L. Gremer, O. Meyer, and R. Huber. 1999. Crystal structure and mechanism of CO dehydrogenase, a molybdo iron-sulfur flavoprotein containing S-selanylcysteine. Proc. Natl Acad. Sci. USA 96:8884–8889.

Dominova, I. N., D. Y. Sorokin, I. V. Kublanov, M. V. Patrushev, and S. V. Toshchakov. 2013. Complete genome sequence of *Salinarchaeum sp.* Strain HArcht-Bsk1T, isolated from hypersaline lake Baskunchak, Russia. Genome Announc. 1: pii: e00505–13.

Douzi, B., A. Filloux, and R. Voulhoux. 2012. On the path to uncover the bacterial type II secretion system. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367:1059–1072.

Eschemann, A., M. Kuhl, and H. Cypionka. 1999. Aerotaxis in *Desulfovibrio*. Environ. Microbiol. 1:489–494.

Fareleira, P., J. Legall, A. V. Xavier, and H. Santos. 1997. Pathways for utilization of carbon reserves in *Desulfovibrio gigas* under fermentative and respiratory conditions. J. Bacteriol. 179:3972–3980.

Felix, R., R. Rodrigues, P. Machado, S. Oliveira, and C. Rodrigues-Pousada. 2006. A chemotaxis operon in the bacterium *Desulfovibrio gigas* is induced under several growth conditions. DNA Seq. 17:56–64.

Fischer-Friedrich, E., G. Meacci, J. Lutkenhaus, H. Chate, and K. Kruse. 2010. Intra- and intercellular fluctuations in Min-protein dynamics decrease with cell length. Proc. Natl Acad. Sci. USA 107:6134–6139.

Frazao, C., G. Silva, C. M. Gomes, P. Matias, R. Coelho, L. Sieker, et al. 2000. Structure of a dioxygen reduction enzyme from *Desulfovibrio gigas*. Nat. Struct. Biol. 7:1041–1045.

Gilmour, C. C., D. A. Elias, A. M. Kucken, S. D. Brown, A. V. Palumbo, C. W. Schadt, et al. 2011. Sulfate-reducing bacterium *Desulfovibrio desulfuricans* ND132 as a model for understanding bacterial mercury methylation. Appl. Environ. Microbiol. 77:3938–3951.

Godde, J. S., and A. Bickerton. 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J. Mol. Evol. 62:718–729.

Greene, E. A., C. Hubert, M. Nemati, G. E. Jenneman, and G. Voordouw. 2003. Nitrite reductase activity of sulphate-reducing bacteria prevents their inhibition by nitrate-reducing, sulphide-oxidizing bacteria. Environ. Microbiol. 5:607–617.

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Haft, D. H., J. Selengut, E. F. Mongodin, and K. E. Nelson. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput. Biol. 1:e60.

Hamilton, W. A. 1985. Sulphate-reducing bacteria and anaerobic corrosion. Annu. Rev. Microbiol. 39: 195–217.

He, Q., K. H. Huang, Z. He, E. J. Alm, M. W. Fields, T. C. Hazen, et al. 2006. Energetic consequences of nitrite stress in *Desulfovibrio vulgaris* Hildenborough, inferred from global transcriptional analysis. Appl. Environ. Microbiol. 72:4370–4381.

Heidelberg, J. F., R. Seshadri, S. A. Haveman, C. L. Hemme, I. T. Paulsen, J. F. Kolonay, et al. 2004. The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. Nat. Biotechnol. 22:554–559.

Helm, R. A., A. G. Lee, H. D. Christman, and S. Maloy. 2003. Genomic rearrangements at rrn operons in *Salmonella*. Genetics 165:951–959.

Hill, N. S., R. Kadoya, D. K. Chattoraj, and P. A. Levin. 2012. Cell size and the initiation of DNA replication in bacteria. PLoS Genet. 8:e1002549.

Hsieh, Y. C., M. Y. Liu, J. le Gall, and C. J. Chen. 2005. Anaerobic purification and crystallization to improve the crystal quality: ferredoxin II from *Desulfovibrio gigas*. Acta Crystallogr. D Biol. Crystallogr. 61: 780–783.

Huang, H., and S. Larter. 2005. Biodegradation of petroleum in subsurface geological reservoirs. Pp. 91–121 *in* B. Olliver and M. Magot, eds. Petroleum microbiology. ASM Press, Washington, DC.

Janssen, A. J., R. Ruitenberg, and C. J. Buisman. 2001. Industrial applications of new sulphur biotechnology. Water Sci. Technol. 44:85–90.

Ji, B., G. Gimenez, V. Barbe, B. Vacherie, Z. Rouy, A. Amrani, et al. 2013. Complete genome sequence of the piezophilic, mesophilic, sulfate-reducing bacterium *Desulfovibrio hydrothermalis* AM13(T.). Genome Announc. 1: pii: e00226–12.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Jorgensen, B. B. 1982. Ecology of the bacteria of the sulphur cycle with special reference to anoxic-oxic interface environments. Philos. Trans. R. Soc. Lond. B Biol. Sci. 298:543–561.

Kakinuma, Y., I. Yamato, and T. Murata. 1999. Structure and function of vacuolar Na+-translocating ATPase in *Enterococcus hirae*. J. Bioenerg. Biomembr. 31:7–14.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Krekeler, C., H. Ziehr, and J. Klein. 1989. Physical methods for characterization of microbial surfaces. Experientia 45:1047–1055.

Lagesen, K., P. Hallin, E. A. Rodland, H. H. Staerfeldt, T. Rognes, and D. W. Ussery. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35:3100–3108.

Le, S. Q., and O. Gascuel. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

LeGALL, J. 1963. A new species of *Desulfovibrio*. J. Bacteriol. 86:1120.

Lenz, M., E. D. Hullebusch, G. Hommes, P. F. Corvini, and P. N. Lens. 2008. Selenate removal in methanogenic and sulfate-reducing upflow anaerobic sludge bed reactors. Water Res. 42:2184–2194.

Li, Q., L. Li, T. Rejtar, D. J. Lessner, B. L. Karger, and J. G. Ferry. 2006. Electron transport in the pathway of acetate conversion to methane in the marine archaeon *Methanosarcina acetivorans*. J. Bacteriol. 188:702–710.

Li, X., M. J. McInerney, D. A. Stahl, and L. R. Krumholz. 2011. Metabolism of H_2 by *Desulfovibrio alaskensis* G20 during syntrophic growth on lactate. Microbiology 157:2912–2921.

Lowe, T. M., and S. R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:955–964.

Lucchetti-Miganeh, C., D. Goudenege, D. Thybert, G. Salbert, and F. Barloy-Hubler. 2011. SORGOdb: Superoxide reductase gene ontology curated DataBase. BMC Microbiol. 11:105.

Makarova, K. S., D. H. Haft, R. Barrangou, S. J. Brouns, E. Charpentier, P. Horvath, et al. 2011. Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. 9:467–477.

Mancini, S., H. K. Abicht, O. V. Karnachuk, and M. Solioz. 2011. Genome sequence of *Desulfovibrio sp.* A2, a highly copper resistant, sulfate-reducing bacterium isolated from effluents of a zinc smelter at the Urals. J. Bacteriol. 193:6793–6794.

Marchler-Bauer, A., C. Zheng, F. Chitsaz, M. K. Derbyshire, L. Y. Geer, R. C. Geer, et al. 2013. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. 41:D348–D352.

Marraffini, L. A., and E. J. Sontheimer. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science 322:1843–1845.

Marshall, W. F., K. D. Young, M. Swaffer, E. Wood, P. Nurse, A. Kimura, et al. 2012. What determines cell size? BMC Biol. 10:101.

Matias, P. M., J. Morais, R. Coelho, M. A. Carrondo, K. Wilson, Z. Dauter, et al. 1996. Cytochrome c_3 from *Desulfovibrio gigas*: crystal structure at 1.8 A resolution and evidence for a specific calcium-binding site. Protein Sci. 5:1342–1354.

Morais-Silva, F. O., C. I. Santos, R. Rodrigues, I. A. Pereira, and C. Rodrigues-Pousada. 2013. Roles of HynAB and Ech, the only two hydrogenases found in the model sulfate reducer *Desulfovibrio gigas*. J. Bacteriol. 195:4753–4760.

Muyzer, G., and A. Stams. 2008. The ecology and biotechnology of sulphate-reducing bacteria. Nat. Rev. Microbiol. 6:441–454.

Nakazawa, H., A. Arakaki, S. Narita-Yamada, I. Yashiro, K. Jinno, N. Aoki, et al. 2009. Whole genome sequence of *Desulfovibrio magneticus* strain RS-1 revealed common gene clusters in magnetotactic bacteria. Genome Res. 19:1801–1808.

Odom, J. M., H. D. Peck, and JR.. 1981. Localization of dehydrogenases, reductases, and electron transfer components in the sulfate-reducing bacterium *Desulfovibrio gigas*. J. Bacteriol. 147:161–169.

Overeijnder, M. L., W. R. Hagen, and P. L. Hagedoorn. 2009. A thermostable hybrid cluster protein from *Pyrococcus furiosus*: effects of the loss of a three helix bundle subdomain. J. Biol. Inorg. Chem. 14:703–710.

Pereira, P. M., M. Teixeira, A. V. Xavier, R. O. Louro, and I. A. Pereira. 2006. The Tmc complex from *Desulfovibrio vulgaris* hildenborough is involved in transmembrane electron transfer from periplasmic hydrogen oxidation. Biochemistry 45:10359–10367.

Pereira, P. M., Q. He, F. M. Valente, A. V. Xavier, J. Zhou, I. A. Pereira, et al. 2008. Energy metabolism in *Desulfovibrio vulgaris* Hildenborough: insights from transcriptome analysis. Antonie Van Leeuwenhoek 93:347–362.

Pereira, I. A., A. R. Ramos, F. Grein, M. C. Marques, S. M. da Silva, and S. S. Venceslau. 2011. A comparative genomic analysis of energy metabolism in sulfate reducing bacteria and archaea. Front Microbiol. 2:69.

Pires, R., A. Lourenço, F. Morais, M. Teixeira, A. Xavier, L. Saraiva, et al. 2003. A novel membrane-bound respiratory complex from *Desulfovibrio desulfuricans* ATCC 27774. Biochim. Biophys. Acta 1605:67–82.

Plugge, C. M., J. C. Scholten, D. E. Culley, L. Nie, F. J. Brockman, and W. Zhang. 2010. Global transcriptomics analysis of the *Desulfovibrio vulgaris* change from syntrophic growth with *Methanosarcina barkeri* to sulfidogenic metabolism. Microbiology 156:2746–2756.

Postgate, J. R., and L. L. Campbell. 1966. Classification of *Desulfovibrio* species, the nonsporulating sulfate-reducing bacteria. Bacteriol Rev 30:732–738.

Raaijmakers, H., S. Macieira, J. M. Dias, S. Teixeira, S. Bursakov, R. Huber, et al. 2002. Gene sequence and the 1.8 A crystal structure of the tungsten-containing formate dehydrogenase from *Desulfovibrio gigas*. Structure 10:1261–1272.

Rabus, R., T. A. Hansen, and F. Widdel. 2006. Dissimilatory sulfate- and sulfur-reducing prokaryotes. Pp. 659–768 in M. Dworkin, S. Falkow, E. Rosenberg, K. H. Sshleifer, E. Stackebrandt, eds. Prokaryotes . 3rd ed. Springer New York, New York, USA.

Ragsdale, S. W., and E. Pierce. 2008. Acetogenesis and the Wood-Ljungdahl pathway of CO(2) fixation. Biochim. Biophys. Acta 1784:1873–1898.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.

Rodionov, D. A., I. Dubchak, A. Arkin, E. Alm, and M. S. Gelfand. 2004. Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria. Genome Biol. 5:R90.

Rodrigues, R., F. M. Valente, I. A. Pereira, S. Oliveira, and C. Rodrigues-Pousada. 2003. A novel membrane-bound Ech [NiFe] hydrogenase in *Desulfovibrio gigas*. Biochem. Biophys. Res. Commun. 306:366–375.

Rodrigues, P. M., A. L. Macedo, B. J. Goodfellow, I. Moura, and J. J. Moura. 2006a. *Desulfovibrio gigas* ferredoxin II: redox structural modulation of the [3Fe-4S] cluster. J. Biol. Inorg. Chem. 11:307–315.

Rodrigues, R., J. B. Vicente, R. Felix, S. Oliveira, M. Teixeira, and C. Rodrigues-Pousada. 2006b. *Desulfovibrio gigas* flavodiiron protein affords protection against nitrosative stress in vivo. J. Bacteriol. 188:2745–2751.

Romao, M. J., M. Archer, I. Moura, J. J. Moura, J. LeGALL, and ENGH, R., SCHNEIDER, M., HOF, P. & HUBER, R.. 1995. Crystal structure of the xanthine oxidase-related aldehyde oxido-reductase from *D. gigas*. Science 270:1170–1176.

Rossi, M., W. B. Pollock, M. W. Reij, R. G. Keon, R. Fu, and G. Voordouw. 1993. The hmc operon of *Desulfovibrio vulgaris* subsp. *vulgaris* Hildenborough encodes a potential transmembrane redox protein complex. J. Bacteriol. 175:4699–4711.

Rothfield, L., A. Taghbalout, and Y. L. Shih. 2005. Spatial control of bacterial division-site placement. Nat. Rev. Microbiol. 3:959–968.

Rousseau, C., M. Gonnet, M. le Romancer, and J. Nicolas. 2009. CRISPI: a CRISPR interactive database. Bioinformatics 25:3317–3318.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, et al. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.

dos Santos, W. G., I. Pacheco, M. Y. Liu, M. Teixeira, A. V. Xavier, and J. Legall. 2000. Purification and characterization of an iron superoxide dismutase and a catalase from the sulfate-reducing bacterium *Desulfovibrio gigas*. J. Bacteriol. 182:796–804.

Sharp, P. M., and W. H. Li. 1987. The codon Adaptation Index- a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Silva, G., S. Oliveira, C. M. Gomes, I. Pacheco, M. Y. Liu, A. V. Xavier, et al. 1999. *Desulfovibrio gigas* neelaredoxin. A novel superoxide dismutase integrated in a putative oxygen sensory operon of an anaerobe. Eur. J. Biochem. 259:235–243.

Silva, G., J. LeGALL, and XAVIER, A. V., TEIXEIRA, M. & RODRIGUES-POUSADA, C.. 2001. Molecular characterization of *Desulfovibrio gigas* neelaredoxin, a protein involved in oxygen detoxification in anaerobes. J. Bacteriol. 183:4413–4420.

Spring, S., M. Visser, M. Lu, A. Copeland, A. Lapidus, S. Lucas, et al. 2012. Complete genome sequence of the sulfate-reducing firmicute *Desulfotomaculum ruminis* type strain (DL(T)). Stand Genomic Sci 7:304–319.

Teske, A., N. B. Ramsing, K. Habicht, M. Fukui, J. Kuver, B. B. Jorgensen, et al. 1998. Sulfate-reducing bacteria and their activities in cyanobacterial mats of solar lake (Sinai, Egypt). Appl. Environ. Microbiol. 64:2943–2951.

Vance, I., and D. R. Thrasher. 2005. Reservoir souring: mechanisms and prevention. Pp. 123–150 in B. Olliver, B., and M. Magot, eds. Petroleum Microbiology. ASM Press, Washington, DC, USA.

Varela-Raposo, A., C. Pimentel, F. Morais-Silva, A. Rezende, J. C. Ruiz, and C. Rodrigues-Pousada. 2013. Role of NorR-like transcriptional regulators under nitrosative stress of the delta-proteobacterium, *Desulfovibrio gigas*. Biochem. Biophys. Res. Commun. 431:590–596.

Venceslau, S. S., R. R. Lino, and I. A. Pereira. 2010. The Qrc membrane complex, related to the alternative complex III, is a menaquinone reductase involved in sulfate respiration. J. Biol. Chem. 285:22774–22783.

Volbeda, A., M. H. Charon, C. Piras, E. C. Hatchikian, M. Frey, and J. C. Fontecilla-Camps. 1995. Crystal structure of the nickel-iron hydrogenase from *Desulfovibrio gigas*. Nature 373:580–587.

Voordouw, G. 2002. Carbon monoxide cycling by *Desulfovibrio vulgaris* Hildenborough. J. Bacteriol. 184:5903–5911.

Walker, C. B., Z. L. He, Z. K. Yang, J. A. Ringbauer, Q. He, J. H. Zhou, et al. 2009. The Electron Transfer System of Syntrophically Grown *Desulfovibrio vulgaris*. J. Bacteriol. 191:5793–5801.

Wang, S., H. Huang, J. Moll, and R. K. Thauer. 2010. NADP+ reduction with reduced ferredoxin and NADP+ reduction with NADH are coupled via an electron-bifurcating enzyme complex in *Clostridium kluyveri*. J. Bacteriol. 192:5115–5123.

Weart, R. B., A. H. Lee, A. C. Chien, D. P. Haeusser, N. S. Hill, and P. A. Levin. 2007. A metabolic sensor governing cell size in bacteria. Cell 130:335–347.

Wolfe, M. T., J. Heo, J. S. Garavelli, and P. W. Ludden. 2002. Hydroxylamine reductase activity of the hybrid cluster protein from *Escherichia coli*. J. Bacteriol. 184:5898–5902.

Zaunmuller, T., D. J. Kelly, F. O. Glockner, and G. Unden. 2006. Succinate dehydrogenase functioning by a reverse redox loop mechanism and fumarate reductase in sulphate-reducing bacteria. Microbiology 152:2443–2453.

Zdobnov, E. M., and R. Apweiler. 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848.

Zerbino, D. R., and E. Birney. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Specific genomic organization of *Desulfovibrio gigas*. (A) Organization of the *hcp* and *hcpR* monocistronic operons in *D.gigas* in comparison with other *Desulfovibrio* species, where: *frdx*, ferredoxin; *a*, Upsa-like protein; *b*, alcohol dehydrogenase; *c*, sensory box histidine kinase; *d*, acpD:acyl carrier protein phosphodiesterase; *e*, putative lipoprotein; *f*, polysaccharide export protein; *g*, cupin 2 conserved barrel domain protein. Gene cluster organization from *D. vulgaris* Hildenborough, *D. alaskensis* G20, and *D. desulfuricans* ATCC27774 were obtained at the DOE Joint Genome

Institute (http://www.jgi.doe.gov/). (B) Organization of operons exclusively present in *D. gigas* genome as compared to other *Desulfovibrio* spp.: (i) aerobic-type carbon monoxide dehydrogenase complex; (ii) vacuolar-type ATP-synthase complex; and (iii) multisubunit Na^+/H^+ antiporter complex. Genes were assigned according to the predicted protein function. The unnamed coding regions are either hypothetical proteins or proteins of unknown function.

Figure S2. Interaction network of *Desulfovibrio gigas* proteins involved in cell size. Purple circles indicate central elements of the network. Yellow circles indicate elements with a fewer number of interactions. Blue lines show protein interactions common to several *D.* genus as retrieved by the STRING database,whereas red lines correspond to *D. gigas* specific interactions.

Table S1. COG functional groups.
Table S2. Codon usage.
Table S3. Transposable elements.
Table S4. Selenocystein-containing proteins.
Table S5. CRISPR proteins.
Table S6. Chemotaxis proteins.
Table S7. Response to oxygen.
Table S8. Nitrogen metabolism.
Table S9. Transcriptional factors sigma 54.
Table S10. Sulfate metabolism.
Table S11. Pentose phosphate pathway.
Table S12. Beta oxidation.
Table S13. Embden-Meyerhof-Parnas pathway.
Table S14: Entner-Doudoroff pathway.
Table S15. TCA cycle.
Table S16. Fumarate metabolism.
Table S17. WoodLjungahl pathway.
Table S18. Alcohol metabolism.
Table S19. Lactate metabolism.
Table S20. Formate metabolism.
Table S21. Oxidation of pyruvate to acetyl-CoA and acetate formation.
Table S22. ATP synthesis.
Table S23. Cytochromes.
Table S24. Hydrogenases.
Table S25. Membranar energy complexes.
Table S26. Nfn complexes.
Table S27. Hdr-like proteins.

SUPPORTING INFORMATION

FIGURES

TABLES

Legends to Figures

Figure S1. Specific genomic organization of *D.gigas*. **(A)** Organization of the *hcp* and *hcpR* monocistronic operons in *D.gigas* in comparison with other *Desulfovibrio* species, where: *frdx*, ferredoxin; *a*, Upsa-like protein; *b*, alcohol dehydrogenase; *c*, sensory box histidine kinase; *d*, acpD:acyl carrier protein phosphodiesterase; *e*, putative lipoprotein; *f*, polysaccharide export protein; *g*, cupin 2 conserved barrel domain protein. Gene cluster organization from *D.vulgaris* Hildenborough, *D.alaskensis* G20 and *D. desulfuricans* ATCC27774 were obtained at the DOE Joint Genome Institute (http://www.jgi.doe.gov/). **(B)** Organization of operons exclusively present in *D.gigas* genome as compared to other *Desulfovibrio* spp.: i) aerobic-type carbon monoxide dehydrogenase complex; ii) vacuolar-type ATP synthase complex; and iii) multisubunit Na+/H+ antiporter complex. Genes were assigned according to the predicted protein function. The unnamed coding regions are either hypothetical proteins or proteins of unknown function.

Figure S2. Interaction network of *D. gigas* proteins involved in cell size. Purple circles indicate central elements of the network. Yellow circles indicate elements with a fewer number of interactions. Blue lines show protein interactions common to several *Desulfovibrio* genus as retrieved by the STRING database .whereas red lines correspond to *D. gigas* specific interactions.

GENERAL GENOME INFORMATION

Table S1. Number of coding regions (CDS´s) associated with the general COG functions

Code	Value	Description
		Information Storage and Processing
K	86	Transcription
J	148	Translation, ribosomal structure and biogenesis
L	104	Replication, recombination and repair
B	1	Chromatin structure and dynamics
		Metabolism
C	184	Energy producing and conversion
F	55	Nucleotide transport and metabolism
H	97	Coenzyme transport and metabolism
Q	24	Secondary metabolites biosynthesis, transport and catabolism
I	36	Lipid transport and metabolism
P	112	Inorganic ion transport and metabolism
G	103	Carbohydrate transport and metabolism
E	208	Amino acid transport and metabolism
		Cellular processes and Signaling
U	23	Intracellular trafficking, secretion and vesicular transport
M	143	Cell wall/membrane/envelope biogenesis
V	37	Defense mechanisms
T	263	Signal transduction mechanisms
N	131	Cell motility
O	91	Posttranslational modification, protein turnover, chaperones
D	32	Cell cycle control, cell division, chromosome partitioning
		Poorly characterized
S	150	Function unknown
R	245	General function predicted only
-	999	No function

Table S2 – Codon usage

Codon	Aminoacid	Fraction	Number of AA	Codon	Aminoacid	Fraction	Number of AA
GCA	Ala	7.1%	9240	CCA	Pro	5.3%	3125
GCC	**Ala**	**66.7%**	**86142**	**CCC**	**Pro**	**48.2%**	**28345**
GCG	Ala	21.1%	27259	CCG	Pro	38.5%	22629
GCT	Ala	5.1%	6604	CCT	Pro	8%	4687
TGC	**Cys**	**86.6%**	**13464**	CAA	Gln	14.4%	6158
TGT	Cys	13.4%	2079	**CAG**	**Gln**	**85.6%**	**36750**
GAC	**Asp**	**65.1%**	**36850**	AGA	Arg	1.9%	1389
GAT	Asp	34.9%	19738	AGG	Arg	4.4%	3241
GAA	**Glu**	**50%**	**32497**	CGA	Arg	3.8%	2825
GAG	**Glu**	**50%**	**32484**	**CGC**	**Arg**	**52.6%**	**38613**
TTC	**Phe**	**67.3%**	**26410**	CGG	Arg	29.1%	21408
TTT	Phe	32.7%	12853	CGT	Arg	8.2%	6018
GGA	Gly	6.9%	5823	AGC	Ser	28.2%	15018
GGC	**Gly**	**67.4%**	**57198**	AGT	Ser	4.3%	2278
GGG	Gly	19.4%	16443	TCA	Ser	2.9%	1545
GGT	Gly	6.3%	5453	**TCC**	**Ser**	**44.6%**	**23742**
CAC	**His**	**60.9%**	**15731**	TCG	Ser	15%	7973
CAT	His	39.1%	10114	TCT	Ser	5%	2660
ATA	Ile	2.8%	1320	ACA	Thr	6.1%	3346
ATC	**Ile**	**75.2%**	**35290**	**ACC**	**Thr**	**60.5%**	**33047**
ATT	Ile	22%	10342	ACG	Thr	29%	15823
AAA	Lys	25.5%	9599	ACT	Thr	4.4%	2405
AAG	**Lys**	**74.5%**	**28081**	GTA	Val	2.6%	2051
CTA	Leu	0.5%	640	GTC	Val	24.2%	19262
CTC	Leu	17.8%	22100	**GTG**	**Val**	**68.7%**	**54613**
CTG	**Leu**	**66.6%**	**82701**	GTT	Val	4.5%	3581
CTT	Leu	5.8%	7257	**TGG**	**Trp**	**100%**	**13642**
TTA	Leu	0.4%	547	**TAC**	**Tyr**	**64.1%**	**15434**
TTG	Leu	8.9%	10925	TAT	Tyr	35.9%	8629
ATG	**Met**	**100%**	**28419**	TAA	*	24.3%	796
AAC	**Asn**	**70.9%**	**18529**	TAG	*	31%	1015
AAT	Asn	29.1%	7592	**TGA**	*	**44.7%**	**1466**

MISCELLANEOUS

Table S3. Transposable elements

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Putative transposase	0060	114		0.372
Putative transposase	0061	118		0.348
Putative transposase	0492	176		0.422
Transposase-like Mu	0917	714		0.660
Transposase, IS4 family protein	1207	466		0.611
Putative transposase	1368	276		0.648
Transposase IS3/IS911 family protein	2017	41		0.698
Integrase catalytic subunit	2025	103		0.486
Integrase catalytic region	2394	732		0.686
Transposase IS4 family protein	2426	50		0.549
Transposase IS4 family protein	2446	353		0.617
Putative transposase	2457	105		0.565
IS4 family transposase	2643	436		0.577
Putative transposase-like protein	3314	159		0.346
ISSoc4,transposase orfA	3315	123		0.399
Putative transposase	3331	299		0.567
Putative transposase-like protein	3366	159		0.342

Table S4. Selenocysteine-containing proteins

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Selenocysteine-specific Translation Elongation Factor	1858	643	selB	0.739
L-seryl-tRNA Selenium Transferase	1861	470	selA	0.744
Conserved hypothetical protein	2096	106		0.706
HesB-like domain-containing protein	2358	106	hesB	0.610
Selenide, water Dikinase	2804	325	selD	0.664
DsrE family protein	3368	106		0.706
Cysteine Desulfurase / Selenocystein Lyase	2344	383	csdA	0.669
Selenium metabolism protein YedF	2272	215	yedF	0.700
Chain A, Tungsten Containing Formate Dehydrogenase	1366	1012	fdh IB	0.752

Table S5. CRISPR- associated proteins

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
CRISPR-associated Protein Cas1, YPEST subtype	1866	326	cas1	0.469
CRISPR-associated Helicase Cas3 family	1867	1109	cas3	0.544
CRISPR-associated Protein, Csy1 family	1868	443	csy1	0.549
CRISPR-associated Protein, Csy2 family	1869	309	csy2	0.586

CRISPR-associated Protein, Csy3 family	1870	345	csy3	0.525
CRISPR-associated Protein, Csy4 family	1871	186	csy4	0.492
Conserved hypothetical protein	2447	337		0.440
CRISPR-associated Helicase	2448	982	cas3	0.462
CRISPR-associated Protein	2449	475	csb2	0.524
CRISPR-associated Protein	2450	404	csb1	0.591

Table S6. Chemotaxis proteins

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
CheB	0024	374	cheB	0.614
Hypothetical protein	0023	648		0.675
CheR	0022	291	cheR	0.662
ParA family protein	0021	261	parA	0.543
CheW	0020	244	cheW	0.686
CheY	0019	372	cheY	0.637
CheA	0018	974	cheA	0.707
Putative methyl-accepting chemotaxis protein	3081	607	mcp	0.696
Chemotaxis protein CheW	3080	174	cheW	0.680
Superoxide dismutase	3082	130	nlr	0.745
Chemotaxis protein CheW	3083	168	cheW	0.703
Methyl-accepting chemotaxis protein	3084	649	mcp	0.684
Anti-sigma-factor antagonist	3079	102		0.614
CheA signal transduction histidine kinase	3078	700		0.708
Methyl-accepting chemotaxis protein	0120	599		0.698
Methyl-accepting chemotaxis sensory transducer protein	0381	672		0.715
Methyl-accepting chemotaxis protein	0398	676		0.693
CheW protein	0399	162		0.663
Hypothetical protein	0400	93		0.674
Chemotaxis protein CheA	0401	691		0.722
Methyl-accepting chemotaxis protein	0422	683		0.728
CheW protein	0423	162		0.698
Methyl-accepting chemotaxis sensory transducer	0498	710		0.744
Methyl-accepting chemotaxis sensory transducer	0662	572		0.695
Response regulator receiver protein	0820	130	cheY	0.602
CheC, inhibitor of MCP methylation	0821	218	cheC	0.721
Chemotaxis protein CheA	0822	101		0.661
Methyl-accepting chemotaxis sensory transducer with Pas/Pac sensor	0980	782		0.734
Chemotaxis sensory transducer protein	0986	813		0.723
Chemotaxis protein	1143	157		0.734
Methyl-accepting chemotaxis sensory transducer	1240	604		0.405
Chemotaxis sensory transducer protein	1382	779		0.693
Methyl-accepting chemotaxis protein	1490	600		0.757
Methyl-accepting chemotaxis protein	1602	544		0.645

Chemotaxis protein CheW	1665	158		0.712
Methyl-accepting chemotaxis sensory transducer with Cache sensor	1693	725		0.753
CheW protein	1942	166		0.677
CheR-type MCP methyltransferase	1943	477		0.663
Chemotaxis protein CheW	1944	212		0.594
Methyl-accepting chemotaxis sensory transducer	1945	583		0.736
CheA signal transduction histidine kinase	1946	776		0.657
Response regulator receiver modulated CheB methylesterase	1947	380		0.635
Multi-sensor hybrid histidine kinase	1948	1042		0.640
Anti-sigma-factor antagonist	1949	103		0.619
CheD family protein	1950	163		0.674
Methyl-accepting chemotaxis sensory transducer	2220	721		0.653
Methyl-accepting chemotaxis sensory transducer with Cache sensor	2263	572		0.666
Multi-sensor hybrid histidine kinase	2603	1195		0.659
CheW protein	2604	158		0.622
Methyl-accepting chemotaxis sensory transducer	2605	596		0.670
Hypotetical protein	2606	90		0.431
CheD-like chemotaxis protein	2639	160		0.546
Methyl-accepting chemotaxis sensory transducer	2675	679		0.642
Methyl-accepting chemotaxis sensory transducer	2707	583		0.682
Methyl-accepting chemotaxis protein	2714	576		0.567
Methyl-accepting chemotaxis sensory transducer	2743	676		0.686
Chemotaxis protei	2780	248		0.779
Chemotaxis protein	3056	167		0.646
Response regulator receiver protein	3057	156		0.599
Methyl-accepting chemotaxis sensory transducer	3240	459		0.604
Chemotaxis protein histidine kinase CheA	3241	701	cheA	0.609
Alkaline phosphatase synthesis transcriptional regulatory protei	3242	121		0.578
Conserved hypothetical protein	3243	100		0.578
Conserved hypothetical protein	3244	493		0.584
Chemotaxis protein CheA	3245	696		0.594
Response regulator receiver protein	3246	121		0.472
Chemotaxis protein methyltransferase CheR	3247	283		0.512
Conserved hypothetical protein	3248	220		0.497
Chemotaxis-specific methylesterase CheB	3249	359		0.590
Chemoreceptor glutamine deamidase CheD	3250	163		0.585
CheW protein	3251	546		0.637
Methyl-accepting chemotaxis protein	3252	650		0.631
Response regulator receiver domain protein	3253	129		0.506
Methyl-accepting chemotaxis protein	3422	879		0.725
Response regulator with CheY-like receiver, AAA-type ATPase, and DNA-binding domains	3423	123		0.664
Chemotaxis sensory transducer	3467	604		0.653
Response regulator receiver protein	3468	120		0.660

CheA signal transduction histidine kinase	3469	710		0.701
Chemotaxis protein methyltransferase	3470	281		0.650
Response regulator receiver modulated CheB methylesterase	3471	360		0.709

GENERAL METABOLISM

Table S7. Response to oxygen

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Putative superoxide dismutase	1536	244		0.717
Neelaredoxin, superoxide reductase / dismutase	3082	130	nlr	0.745
Catalase	2858	502	cat	0.788
Bacterioferritin	2857	177	bfr	0.785
Rubredoxin-like protein	1167	71	rub2	0.732
Chain A, Chain B, Rubredoxin-oxygen Oxidoreductase	1622	402	roo	0.802
Chain A rubredoxin	1624	52	rd	0.818
Cytochrome bd Quinol Oxidase, subunit I	1252	443	cydA	0.716
Cytochrome bd Quinol Oxidase, subunit II	1253	336	cydB	0.788
Desulforedoxin	3485	37	dsr	0.829
Peroxiredoxin	3518	222	prxU	0.738
Rubrerythrin	0750	165	rbr	0.761
Rubrerythrin	1055	156	rbr	0.684
Rubrerythrin	1714	162		0.710
Coenzyme F390 synthetase	2876	421		0.833
Coenzyme F390 synthetase	3140	441		0.784
Coenzyme F390 synthetase-like	0910	433		0.647

Table S8. Nitrogen metabolism

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Putative Nitrogenase Cofactor Biosynthesis Protein NifB	1271	423		0.810
Putative Nitrogenase	1272	471		0.770
Putative Nitrogenase MoFe Cofactor Biosynthesis Protein	1273	475		0.796
Putative Nitrogenase Cofactor Biosynthesis Protein NifB	1275	403		0.764
Putative Nitrogenase Molybdenum-iron Protein, beta chain	1276	461		0.805
Putative Nitrogenase Molybdenum-iron Protein, subunit alpha	1277	555		0.824
Putative Nitrogen Regulatory Protein P-II	1278	126		0.746
Putative Nitrogen Regulatory Protein P-II	1279	118		0.693
Nitrogenase Reductase	1280	274	nifH	0.767

Putative Nitrogen Regulatory Protein P-II	2553	112	glnB-1	0.790
Putative Ammonium Transporter	2554	402	amt	0.654
Putative Nitrate Reductase	0241	698		0.732
Putative Nitrate Reductase	1101	640		0.724
Putative Nitrate Reductase	1195	705		0.710
Putative Transcriptional Regulator, NifA, Fis Family	1208	515	norR2L	0.719
Putative Fis family NifA subfamily Transcriptional Regulator	0080	525	norR1L	0.743
Putative Cytochrome c Nitrite Reductase, small subunit	1513	143	nrfH	0.754
Putative Nitrite Reductase	1514	488	nrfA	0.780
Putative Hydroxylamine Reductase	1496	545	hcp	0.817
Putative cAMP-binding protein	1495	224	hcpR	0.748
Putative Nitroreductase	0813	174		0.720
Putative Nitroreductase	1466	305		0.739
Putative Nitroreductase	2864	169		0.652
Putative Carbamoyl-phosphate Synthase, large subunit	0189	1082		0.816
Putative Carbamoyl-phosphate Synthase, small subunit	0743	383	carA	0.782
Putative Ornithine Carbamoyltransferase	2162	301	argF	0.717
Putative Argininosuccinate Synthase	2161	403	argG	0.820
Putative Argininosuccinate Lyase	2160	462		0.732
Putative Glutamate Synthase	2742	507		0.740
Putative Glutamate Synthase (NADPH), homotetrameric	1689	481	nfnB	0.769
Putative Glutamine Synthetase, type I	1150	448		0.764

Table S9. Transcriptional factors Sigma 54

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Putative Sigma-54 dependent Transcriptional Regulator/Response Regulator	0875	458		0.766
Putative two component, Sigma54 specific, Transcriptional Regulator, Fis family	0999	478		0.734
Putative Fis family two component, Sigma-54 specific, Transcriptional Regulator	1046	453		0.796
Putative two component, Sigma54 specific, Fis family Transcriptional Regulator	1069	486		0.753
Putative Sigma-54 Factor Interaction domain-containing Protein	1255	468		0.714
Putative two component, Sigma54 specific, Transcriptional Regulator	1424	507		0.664
Putative Fis family Sigma-54 specific activator	1482	338		0.663
Putative two component Sigma-54 specific Transcriptional Regulator	1580	470		0.777
Putative Fis family two component Sigma-54 specific Transcriptional Regulator	1653	473		0.669
Putative Fis family two component Sigma-54 specific	1941	479		0.726

Transcriptional Regulator				
Putative ECF subfamily RNA Polymerase Sigma-24	2673	202		0.633
Putative Sigma 54 interacting domain Protein	3035	844		0.512
Putative PAS modulated Sigma54 specific Transcriptional Regulator, Fis family	1702	441		0.632

Table S10. Sulfate metabolism

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Adenylylsulfate Reductase B subunit	0458	167	aprB	0.757
Adenylylsulfate Reductase A subunit	0457	666	aprA	0.734
Adenylylsulfate Kinase	1831	196	cycC	0.646
Desulforedoxin	3485	37	dsr	0.829
Cobyrinic Acid A,C-diamide Synthase	0687	487		0.679
Dissimilatory Sulfite Reductase B	0688	75	dsrD	0.634
Dissimilatory Sulfite Reductase I (Dsri) B subunit	0689	386	dsrB	0.738
Dissimilatory Sulfite Reductase I (Dsri) A subunit	0691	437	dsrA	0.723
Dissimilatory Sulfite Reductase, gamma subunit	1681	105	dsrC	0.735
DsrE family protein	1238	118		0.759
Nitrite and Sulfite Reductase, 4Fe-4S region	3367	217		0.772
DsrE family protein	3368	106		0.706
Phosphoadenosine Phosphosulfate Reductase	3007	273	paps	0.576
Sulfate Adenylyltransferase	0460	426	sat	0.778
DnaJ-like Protein	0854	132		0.752
Protein of unknown function	0855	211		0.670
Mammalian cell entry domain-containing protein	0856	310		0.671
Sulfate-transporting ATPase	0857	259		0.711
Protein of unknown function	0858	250		0.678
Putative Sulfate Transport Protein CysZ	0923	214		0.740
Putative Sulfate Transport Protein CysZ	2397	205		0.599
Sulfate Permease family Protein	0373	561	sulP	0.702

CENTRAL AND ENERGY METABOLISM

Table S11. Pentose Phosphate Pathway

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Transketolase	1348	666	tkt	0.761
Glucose-6-Phosphate Isomerase	2341	552	gpi	0.708
Ribulose-Phosphate 3-Epimerase	1351	226	rpe	0.740
Translaldolase	2275	220	tal	0.666
6-Phosphogluconate Dehydrogenase, Decarboxylating	1384	301	gnd	0.717
Glucose-6-Phosphate 1-Dehydrogenase	1385	523	zwf	0.679
6-Phosphogluconolactonase	1386	242	pgl	0.578
Ribose-5-Phosphate Isomerase A	1387	236	rpi	0.662

Table S12. Beta Oxidation

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
N-acetyltransferase GCN5	2273	161		0.518
Acyl-CoA Dehydrogenase domain-containing protein	1478	689		0.756
Electron-transferring-flavoprotein Dehydrogenase	1479	618		0.762
Electron Transfer Flavoprotein, Beta subunit	1480	268	etfB	0.711
Electron Transfer Flavoprotein, Alpha subunit	1481	339	etfA	0.717

Table S13. Embden-Meyerhof-Parnas Pathway

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Enolase (phosphopyruvate hydratase)	0704	441	eno	0.822
Fructose-bisphosphate Aldolase	0591	267	fbaB	0.702
Putative Phospho-2-dehydro-3-deoxyheptonate Aldolase	0592	267	fbaB2	0.722
Fructose-bisphosphate Aldolase	1026	257	fbaB1	0.724
Fructose-1,6-bisphosphatase	2524	349	fbp	0.696
Type I Glyceraldehyde-3-phosphate Dehydrogenase	2188	333	gap	0.766
Glyceraldehyde-3-phosphate Dehydrogenase, type I	2546	338	gap1	0.755
Glucokinase	1383	327	gck	0.654
Glycogen Phosphorylase	2153	855	glgP	0.731
Phosphoglycerate Mutase 1 family	1753	248	gpmA	0.774
2,3-Bisphosphoglycerate-independent Phosphoglycerate Mutase	3489	532	gpmB	0.714
PEP Synthase	2938	848		0.791
Diphosphate--fructose-6-phosphate 1-Phosphotransferase	2195	456	pfkA	0.747
Phosphoglycerate Kinase	1347	391	pgk	0.784
Phosphoglucomutase, Alpha-D-glucose Phosphate-	0235	549	pgm	0.785

specific				
PTS System Mannose/fructose/sorbose IID component family protein	3110	462		0.711
Phosphocarrier protein HPr	3111	162	ptsH	0.758
Phosphoenolpyruvate-protein Phosphotransferase	3112	594	ptsA	0.749
PTS System Sorbose subfamily transporter subunit IIB	2413	156		0.741
PTS System Fructose IIA component family protein	2414	149		0.655
PTS IIA-like Nitrogen-regulatory Protein PtsN	2416	149	ptsN	0.640
Pyruvate Kinase	0179	483	pyk	0.726
Triose-phosphate Isomerase	0332	249	tpi	0.569

Table S14. Entner-Doudoroff Pathway

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Dihydroxy-Acid Dehydratase	0845	555	edd	0.806
Glucose-6-phosphate 1-Dehydrogenase	1385	523	zwf	0.679
Glucokinase	1383	327	gck	0.654

Table S15. TCA Cycle

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Aconitate Hydratase	0463	651	aco	0.788
Fumarate Hydratase	1571	278	fumA	0.821
Fe-S Type, Tartrate/fumarate subfamily Hydro-lyase subunit alpha	1573	181	fumC	0.790
Citrate Synthase I	1379	436	gltA	0.737
Isocitrate Dehydrogenase	1851	382	icd	0.806
Malic Protein NAD-binding protein	1574	437	mdh	0.803

Table S16. Fumarate Metabolism

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Fumarate Reductase Respiratory Complex	1568	218	fdrC	0.643
Fumarate Reductase, Flavoprotein subunit	1569	627	fdrA	0.810
Fumarate Reductase, Iron-sulfur subunit	1570	264	fdrB	0.744
Fumarate reductase Respiratory Complex, transmembrane subunit	0826	224	fdrCII	0.686
Fumarate Reductase, Flavoprotein subunit	0827	615	fdrAII	0.811
Fumarate Reductase, Iron-sulfur subunit	0828	255	fdrBII	0.792

Table S17. WoodLjungdahl Pathway

Encoded Protein	*D.gigas*	# AA	Gene	CAI

	ID		symbol	Index
Acyl-CoA Synthetase (NDP forming)	0571	905	acsA	0.751
Homocysteine S-Methyltransferase	3461	816	acsE	0.709
Aerobic-type Carbon Monoxide Dehydrogenase, small subunit CoxS/CutS-like protein	3143	228	coxS	0.701
Aerobic-type Carbon Monoxide Dehydrogenase, large subunit CoxL/CutL-like protein	3144	782	coxL	0.737
Aerobic-type Carbon Monoxide Dehydrogenase, middle subunit CoxM/CutM-like protein	3145	325	coxM	0.674
Bifunctional 5,10-Methylene-tetrahydrofolate Dehydrogenase/ 5,10-Methylene-tetrahydrofolate Cyclohydrolase	0702	295	folD	0.726
5,10-Methylenetetrahydrofolate Reductase	0415	304	metF	0.736
Cobalamin B12-Binding domain Protein	2600	229	mtsB	0.599

Table S18. Alcohol metabolism

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Iron-containing Alcohol Dehydrogenase	0712	400		0.695
Iron-containing Alcohol Dehydrogenase	1044	380		0.729
Iron-containing Alcohol Dehydrogenase	1047	393		0.749
Zinc-containing Alcohol Dehydrogenase	2585	323		0.657
Zinc-containing Alcohol Dehydrogenase	3525	425		0.688
Aldehyde Dehydrogenase	0746	477	ald	0.772

Table S19. Lactate Metabolism

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
D-Lactate Dehydrogenase	1422	460	ldh	0.833
Glycolate Oxidase subunit GlcD	3110	462		0.711
FMN-dependent Alpha-hydroxy Acid Dehydrogenase	1415	345		0.724
Conserved hypothetical protein	1416	721		0.750
Lactate Utilization Protein B/C	1417	208		0.741
Putative L-Lactate Transport	1423	566		0.749

Table S20. Formate Metabolism

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Chain B, Tungsten Containing Formate Dehydrogenase	1364	245	fdhB	0.745
Chain A, Tungsten Containing Formate Dehydrogenase	1366	1012	fdh A	0.752
Formate Dehydrogenase, alpha subunit	3334	1009	fdnG	0.771
Putative Fe-S-cluster-containing Hydrogenase component protein	3335	247		0.701

Formate Dehydrogenase accessory protein	0759	315	fdhE	0.739
Formate Dehydrogenase subunit FdhD	0760	236	fdhD	0.667
Putative Formate Dehydrogenase, formation protein FdhE1	3336	303	fdhEI	0.636

Table S21. Oxidation of pyruvate to acetyl-CoA and acetate formation

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Aldehyde Ferredoxin Oxidoreductase	1447	561		0.633
Aldehyde Ferredoxin Oxidoreductase	3127	579		0.784
Ferredoxin I	1020	62		0.741
Ferredoxin II	0418	59		0.700
Conserved hypothetical protein	2139	86		0.749
Ferrodoxin-like protein	0869	81	vorC	0.720
2-Ketoisovalerate Ferredoxin Reductase	0870	353	vorA	0.768
Thiamine pyrophosphate binding domain-containing protein	0871	270	vorB	0.752
2-Oxoacid:ferredoxin Oxidoreductase, gamma subunit	0872	187	vorG	0.784
Indolepyruvate Ferredoxin Oxidoreductase	0081	201		0.657
Indolepyruvate Ferredoxin Oxidoreductase subunit alpha	0082	632	iorA	0.707
Pyruvate-Ferredoxin Oxidoreductase	0996	1213	poR	0.805
Pyruvate Ferredoxin/Flavodoxin Oxidoreductase subunit beta	1712	283	porB	0.679
Pyruvate Flavodoxin/Ferredoxin Oxidoreductase domain protein	1713	563	porA	0.642
Biotin/Acetyl-CoA-Carboxylase Ligase	3401	326	pycA	0.556
Pyruvate Carboxylase	3402	1268	pycB	0.754
Pyruvate, water Dikinase., Phosphoenolpyruvate--protein Phosphotransferase	2250	1208	ppdk	0.789
Pyruvate, water Dikinase	3042	883		0.721
Phosphoenolpyruvate Synthase/Pyruvate Phosphate Dikinase	3046	882		0.764
Pyruvate, water Dikinase	2942	841		0.760

Table S22. ATP synthesis

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
ATP synthase F0F1 subunit epsilon	0648	133	atpC	0.616
ATP synthase F1, beta subunit	0649	471	atpD	0.728
F0F1 ATP synthase subunit gamma	0650	293	atpG	0.602
F0F1 ATP synthase subunit alpha	0651	502	atpA	0.752
ATP synthase F0F1 subunit delta	0652	183	atpH	0.591
H+transporting two-sector ATPase subunit B/B'	0653	189	atpF	0.604
H+transporting two-sector ATPase subunit B/B'	0654	138	atpF1	0.548
ATP synthase protein I	1499	248	atpI	0.499

ATP synthase I	1500	156	uncI	0.521
ATP synthase F0 subunit alpha	1501	234	atpB	0.671
ATP synthase F0, C subunit	1502	110	atpE	0.677
V-type ATP synthase subunit K	3061	162		0.698
V-type ATPase 116 kDa subunit	3062	622		0.672
H(+)-transporting ATP synthase, vacuolar type, subunit D	3063	205		0.721
V-type ATP synthase subunit B	3064	430		0.776
V-type ATP synthase subunit A	3065	583		0.768
Two-sector ATPase, V(1) subunit E	3067	214		0.639

Table S23. Cytochromes

Encoded Protein	*D.gigas ID*	# AA	Gene symbol	CAI Index
Cytochrome bd Quinol Oxidase subunit I	1252	443	cydA	0.716
Cytochrome bd Quinol Oxidase subunit II	1253	336	cydB	0.788
Cytochrome c class III	0326	133		0.492
Cytochrome c3	0144	137	cyc	0.767
Di-Tetraheme Cytochrome C3	1464	112		0.834
Cytochrome c554	0380	157		0.689
Respiratory Nitrite Reductase specific menaquinol--cytochrome-c reductase (NrfH) precursor	1513	143	nrfH	0.754
Nitrite Reductase (cytochrome, ammonia-forming)	1514	488	nrfA	0.780
Conserved protein of unknown function	2209	47		0.619
Cytochrome c assembly protein	2210	225	ccmC	0.636
ccmB family protein	2211	224	ccmB	0.554
Heme exporter protein CcmA	2212	226	ccmA	0.692
Cytochrome C assembly protein	2213	635	ccmF	0.714
Conserved hypothetical protein	2214	458		0.643
Cytochrome c-type biogenesis protein CcmE	2216	140	ccmE	0.682
Protein of unknown function	2217	360		0.700

Table S24. Hydrogenases

Encoded Protein	D.gigas ID	# AA	Gene symbol	CAI Index
hynD	2259	82	hynD	0.695
hynC	2260	165	hynC	0.756
hynB	2261	551	hynB	0.802
hynA	2262	288	hynA	0.717
echA	0034	647	echA	0.642
echB	0035	284	echB	0.716
echC	0036	147	echC	0.727
echD	0037	125	echD	0.704
echE	0038	358	echE	0.775

echF	0039	123	echF	0.795
(NiFe) Hydrogenase maturation protein HypF	0896	818	hypF	0.679
Hydrogenase expression/formation protein HypD	1098	372	hypD	0.764
Hydrogenase expression/formation protein HypE	1099	340	hypE	0.706
Hydrogenase accessory protein HypB	2238	219	hypB	0.729
HypA	2239	121	hypA	0.760

ENERGY CONSERVATION

Table S25. Membranar Energy Complexes

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Hdr-like Menaquinol Oxidoreductase Cytochrome b-like subunit	2334	338	dsrM	0.658
Cytoplasmic, binds 2 (4Fe-4S)	2335	542	dsrK	0.758
Periplasmic (Sec) Triheme Cytochrome c	2336	127	dsrJ	0.731
Periplasmic (Tat), binds 2(4Fe-4S)	2337	268	dsrO	0.798
Polysulphide Reductase NrfD	2338	386	dsrP	0.756
Sixteen Heme Cytochrome	0715	559	hmcA	0.697
4Fe-4S Ferredoxin	0716	364	hmcB	0.735
HMC redox complex, integral membrane protein HmcC	0717	416	hmcC	0.796
protein HmcD	0718	48	hmcD	0.622
HMC redox complex, integral membrane protein HmcE	0719	225	hmcE	0.744
protein HmcF	0720	470	hmcF	0.796
NAD(P)H-quinone Oxidoreductase subunit 3	0242	125	nuoA	0.568
NADH-quinone Oxidoreductase subunit B	0243	192	nuoB	0.670
NADH:ubiquinone Oxidoreductase 27 kD subunit	0244	568	nuoC/D	0.764
NADH Dehydrogenase (quinone)	0245	328	nuoH	0.670
NADH:ubiquinone Oxidoreductase chain I-like protein	0246	218	nuoI	0.681
NADH-ubiquinone/plastoquinone Oxidoreductase chain 6	0247	172	nuoJ	0.624
NADH-quinone Oxidoreductase subunit K 1	0248	113	nuoK	0.570
NADH/Ubiquinone/plastoquinone (Complex I)	0249	492	nuoL	0.747
hypothetical protein B193_0141	0250	90		0.707
Monovalent cation/H+ Antiporter subunit D	0251	613	nuoN	0.760
Proton-translocating NADH-quinone Oxidoreductase subunit M	0253	521	nuoM	0.737
NADH Dehydrogenase (quinone)	0254	475		0.745
4Fe-4S Ferredoxin	0255	169		0.816
Permease	0256	378	nuoP	0.750
Conserved hypothetical protein	0257	219		0.739
Multicomponent Na+/H+ Antiporter subunit E	2843	164	mnhE	0.705
Multiple Resistance and pH Regulation protein F	2844	100	mnhF	0.660
Multicomponent Na+/H+ Antiporter subunit G	2845	122	mnhG	0.572

Conserved hypothetical protein	2846	81		0.602
Putative Monovalent cation/H+ Antiporter subunit B	0247	266	mnhB	0.588
Na(+)/H(+) Anitporter subunit MhnC	0248	129	mnhC	0.583
NADH/ubiquinone/plastoquinone	2849	465		0.662
NADH Dehydrogenase (quinone)	2850	513	mnhA	0.561
Conserved hypothetical protein	2851	81		0.492
Putative Monovalent cation/H+ Antiporter subunit D	2852	598	mnhD	0.654
Crp family Transcriptional Regulator	2853	164		0.615
Hydrogenase, b-type Cytochrome subunit	0374	202	ohcC	0.724
Cytochrome c family protein	0375	545	ohcA	0.735
Iron-sulfur Cluster-binding Protein	0376	157		0.606
4Fe-4S Ferredoxin	0377	323	ohcB	0.666
QmoD protein	0453	246	qmoD	0.674
Heterodisulfide Reductase	0454	393	qmoC	0.734
Quinone-interacting Membrane-bound Oxidoreductase	0455	768	qmoB	0.768
Heterodisulfide Reductase	0456	411	qmoA	0.691
Quinone-interacting Membrane-bound Oxidoreductase complex subunit C	2765	402		0.694
Conserved hypothetical protein	2766	228		0.762
Polysulfide Reductase NrfD	0130	411	qrcD	0.774
Molybdopterin Oxidoreductase, Iron-sulfur Cluster-binding subunit	0131	266	qrcC	0.691
Molybdopterin Oxidoreductase	0132	693	qrcB	0.706
Cytochrome C	0133	212	qrcA	0.623
CheY-like receiver, AAA-type ATPase, and DNA-binding domain containing response regulator	1424	507		0.664
ApbE family Lipoprotein	1425	344	rnfF	0.750
4Fe-4S Ferredoxin	1426	293	rnfB	0.772
RnfABCDGE type electron transport complex subunit A	1427	191	rnfA	0.658
RnfABCDGE type electron transport complex subunit E	1428	239	rnfE	0.657
RnfABCDGE type electron transport complex subunit G	1429	193	rnfG	0.773
RnfABCDGE type electron transport complex subunit D	1430	318	rnfD	0.671
4Fe-4S Ferredoxin	1431	398	rnfC	0.730
Cytochrome c family protein	1432	245		0.710
Transmembrane complex, Tetraheme Cytochrome c3	1698	137	tmcA	0.760
Iron-sulfur binding Protein	1699	444	tmcB	0.770
Conserved hypothetical protein	1700	219	tmcC	0.647
TmC complex protein, subunit D	1701	419	tmcD	0.695

Table S26. Nfn complex

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
FAD-dependent Pyridine Nucleotide-disulfide	1688	458		0.650

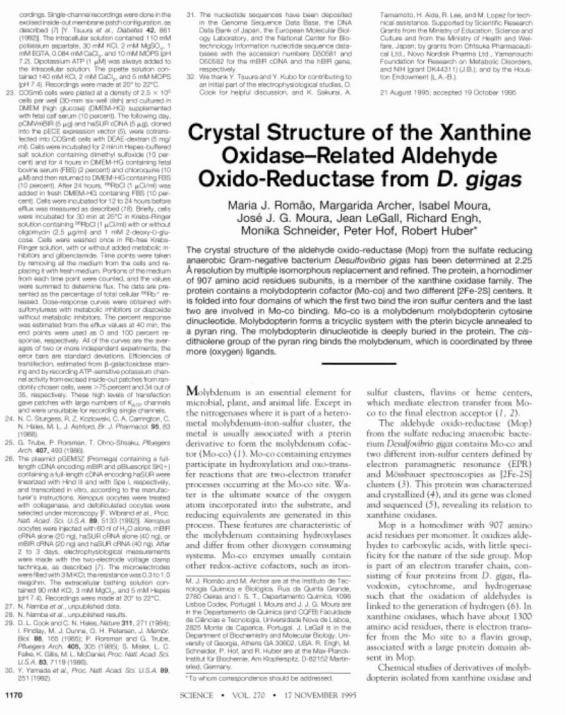
Oxidoreductase				
Oxidoreductase	1689	481	nfnB	0.769
Ferredoxin-NADP(+) Reductase subunit alpha	1690	282	nfnA	0.736
Oxidoreductase	1577	475		0.768
Ferredoxin-NADP Reductase	1578	261		0.717

Table S27. Hdr-like proteins

Encoded Protein	*D.gigas* ID	# AA	Gene symbol	CAI Index
Heterodisulfide Reductase subunit C	1048	197	hdrC	0.711
CoB--CoM Heterodisulfide Reductase	1049	299	hdrB	0.769
Heterodisulfide Reductase, subunit A	1050	669	hdrA	0.786
Methyl-viologen-reducing Hydrogenase delta subunit	1051	149	floxD	0.765
Coenzyme F420 Hydrogenase/Dehydrogenase, beta subunit	1052	321	floxC	0.707
Hydrogenase, putative	1053	403	floxB	0.742
Dihydroorotate Dehydrogenase, electron transfer subunit protein	1054	277	floxA	0.778
FAD linked Oxidase domain-containing Protein	1343	1199	hdrD	0.749
Iron-sulfur cluster-binding protein	1416	721		0.750
Iron-sulfur cluster-binding protein	1421	426		0.798
Fe-S Oxidoreductase	3109	379		0.677
Aldehyde Dehydrogenase, iron-sulfur subunit	0904	762		0.673

IV.2 – The MOP operon

In 1994, our team cloned and published the sequence of the gene from *D. gigas* coding for molybdenum-containing aldehyde oxido-reductase (Thoemes *et al*, 1994). This was the team's first adventure into the genetics of this bacteria because, by that time, this sequence was needed for solving the 3D structure of that protein, which resulted in a high-impact factor publication in Science by Romão et al, 1995 (fig.63). These two publications have been referred 47 times and 449 times respectively.



cordings. Single-channel recordings were done in the excised inside-out membrane patch configuration, as described (7) [Y. Tsuura *et al., Diabetes* **42**, 861 (1993)]. The intracellular solution contained 110 mM potassium aspartate, 30 mM KCl, 2 mM $MgSO_4$, 1 mM EGTA, 0.084 mM $CaCl_2$, and 10 mM MOPS (pH 7.2). Dipotassium ATP (1 μM) was always added to the intracellular solution. The pipette solution contained 140 mM KCl, 2 mM $CaCl_2$, and 5 mM MOPS (pH 7.4). Recordings were made at 20° to 22°C.

23. COSm6 cells were plated at a density of 2.5×10^5 cells per well (30-mm six-well dish) and cultured in DMEM (high glucose) (DMEM-HG) supplemented with fetal calf serum (10 percent). The following day, pCMVmBIR (5 μg) and hSUR cDNA (5 μg), cloned into the pECE expression vector (5), were cotransfected into COSm6 cells with DEAE-dextran (5 mg/ml). Cells were incubated for 2 min in Hepes-buffered salt solution containing dimethyl sulfoxide (10 percent) and for 4 hours in DMEM-HG containing fetal bovine serum (FBS) (2 percent) and chloroquine (10 μM) and then returned to DMEM-HG containing FBS (10 percent). After 24 hours, $^{86}RbCl$ (1 μCi/ml) was added in fresh DMEM-HG containing FBS (10 percent). Cells were incubated for 12 to 24 hours before efflux was measured as described (18). Briefly, cells were incubated for 30 min at 25°C in Krebs-Ringer solution containing $^{86}RbCl$ (1 μCi/ml) with or without oligomycin (2.5 μg/ml) and 1 mM 2-deoxy-D-glucose. Cells were washed once in Rb-free Krebs-Ringer solution, with or without added metabolic inhibitors and glibenclamide. Time points were taken by removing all the medium from the cells and replacing it with fresh medium. Portions of the medium from each time point were counted, and the values were summed to determine flux. The data are presented as the percentage of total cellular $^{86}Rb^+$ released. Dose-response curves were obtained with sulfonylureas with metabolic inhibitors or diazoxide without metabolic inhibitors. The percent response was estimated from the efflux values at 40 min; the end points were used as 0 and 100 percent response, respectively. All of the curves are the averages of two or more independent experiments; the error bars are standard deviations. Efficiencies of transfection, estimated from β-galactosidase staining and by recording ATP-sensitive potassium channel activity from excised inside-out patches from randomly chosen cells, were >75 percent and 34 out of 35, respectively. These high levels of transfection gave patches with large numbers of K_{ATP} channels and were unsuitable for recording single channels.

24. N. C. Sturgess, R. Z. Kozlowski, C. A. Carrington, C. N. Hales, M. L. J. Ashford, *Br. J. Pharmacol.* **95**, 83 (1988).
25. G. Trube, P. Rorsman, T. Ohno-Shosaku, *Pflügers Arch.* **407**, 493 (1986).
26. The plasmid pGEM3Z (Promega) containing a full-length cDNA encoding mBIR and pBluescript SK(+) containing a full-length cDNA encoding haSUR were linearized with Hind III and with Sph I, respectively, and transcribed in vitro, according to the manufacturer's instructions. *Xenopus* oocytes were treated with collagenase, and defolliculated oocytes were selected under microscopy [F. Wibrand *et al., Proc. Natl. Acad. Sci. U.S.A.* **89**, 5133 (1992)]. *Xenopus* oocytes were injected with 50 nl of H_2O alone, mBIR cRNA alone (20 ng), haSUR cRNA alone (40 ng), or mBIR cRNA (20 ng) and haSUR cRNA (40 ng). After 2 to 3 days, electrophysiological measurements were made with the two-electrode voltage clamp technique, as described (7). The microelectrodes were filled with 3 M KCl; the resistance was 0.3 to 1.0 megohm. The extracellular bathing solution contained 90 mM KCl, 3 mM $MgCl_2$, and 5 mM Hepes (pH 7.4). Recordings were made at 20° to 22°C.
27. N. Nanba *et al.*, unpublished data.
28. N. Nanba *et al.*, unpublished results.
29. D. L. Cook and C. N. Hales, *Nature* **311**, 271 (1984); I. Findlay, M. J. Dunne, O. H. Petersen, J. Membr. Biol. **86**, 165 (1985); P. Rorsman and G. Trube, *Pflügers Arch.* **405**, 305 (1985); S. Misler, L. C. Falke, K. Gillis, M. L. McDaniel, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7119 (1986).
30. Y. Yamada *et al., Proc. Natl. Acad. Sci. U.S.A.* **89**, 251 (1992).

31. The nucleotide sequences have been deposited in the Genome Sequence Data Base, the DNA Data Bank of Japan, the European Molecular Biology Laboratory, and the National Center for Biotechnology Information nucleotide sequence databases with the accession numbers D50581 and D50582 for the mBIR cDNA and the hBIR gene, respectively.

32. We thank Y. Tsuura and Y. Kubo for contributing to an initial part of the electrophysiological studies, D. Cook for helpful discussion, and K. Sakura, A. Yamamoto, H. Aida, R. Lee, and M. Lopez for technical assistance. Supported by Scientific Research Grants from the Ministry of Education, Science and Culture and from the Ministry of Health and Welfare, Japan, by grants from Ohtsuka Pharmaceutical Ltd., Novo Nordisk Pharma Ltd., Yamanouchi Foundation for Research on Metabolic Disorders, and NIH (grant DK44311) (J.B.); and by the Houston Endowment (L.A.-B.).

21 August 1995; accepted 19 October 1995

Crystal Structure of the Xanthine Oxidase–Related Aldehyde Oxido-Reductase from *D. gigas*

Maria J. Romão, Margarida Archer, Isabel Moura, José J. G. Moura, Jean LeGall, Richard Engh, Monika Schneider, Peter Hof, Robert Huber*

The crystal structure of the aldehyde oxido-reductase (Mop) from the sulfate reducing anaerobic Gram-negative bacterium *Desulfovibrio gigas* has been determined at 2.25 Å resolution by multiple isomorphous replacement and refined. The protein, a homodimer of 907 amino acid residues subunits, is a member of the xanthine oxidase family. The protein contains a molybdopterin cofactor (Mo-co) and two different [2Fe-2S] centers. It is folded into four domains of which the first two bind the iron sulfur centers and the last two are involved in Mo-co binding. Mo-co is a molybdenum molybdopterin cytosine dinucleotide. Molybdopterin forms a tricyclic system with the pterin bicycle annealed to a pyran ring. The molybdopterin dinucleotide is deeply buried in the protein. The *cis*-dithiolene group of the pyran ring binds the molybdenum, which is coordinated by three more (oxygen) ligands.

Molybdenum is an essential element for microbial, plant, and animal life. Except in the nitrogenases where it is part of a heterometal molybdenum-iron-sulfur cluster, the metal is usually associated with a pterin derivative to form the molybdenum cofactor (Mo-co) (*1*). Mo-co containing enzymes participate in hydroxylation and oxo-transfer reactions that are two-electron transfer processes occurring at the Mo-co site. Water is the ultimate source of the oxygen atom incorporated into the substrate, and reducing equivalents are generated in this process. These features are characteristic of the molybdenum containing hydroxylases that differ from other dioxygen consuming systems. Mo-co enzymes usually contain other redox-active cofactors, such as iron-sulfur clusters, flavins or heme centers, which mediate electron transfer from Mo-co to the final electron acceptor (*1*, *2*).

The aldehyde oxido-reductase (Mop) from the sulfate reducing anaerobic bacterium *Desulfovibrio gigas* contains Mo-co and two different iron-sulfur centers defined by electron paramagnetic resonance (EPR) and Mössbauer spectroscopies as [2Fe-2S] clusters (*3*). This protein was characterized and crystallized (*4*), and its gene was cloned and sequenced (*5*), revealing its relation to xanthine oxidases.

Mop is a homodimer with 907 amino acid residues per monomer. It oxidizes aldehydes to carboxylic acids, with little specificity for the nature of the side group. Mop is part of an electron transfer chain, consisting of four proteins from *D. gigas*, flavodoxin, cytochrome, and hydrogenase such that the oxidation of aldehydes is linked to the generation of hydrogen (*6*). In xanthine oxidases, which have about 1300 amino acid residues, there is electron transfer from the Mo site to a flavin group, associated with a large protein domain absent in Mop.

Chemical studies of derivatives of molybdopterin isolated from xanthine oxidase and

M. J. Romão and M. Archer are at the Instituto de Tecnologia Química e Biológica, Rua da Quinta Grande, 2780 Oeiras and I. S. T., Departamento Química, 1096 Lisboa Codex, Portugal. I. Moura and J. J. G. Moura are in the Departamento de Química (and CQFB) Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2825 Monte da Caparica, Portugal. J. LeGall is in the Department of Biochemistry and Molecular Biology, University of Georgia, Athens GA 30602, USA. R. Engh, M. Schneider, P. Hof, and R. Huber are at the Max-Planck-Institut für Biochemie, Am Klopferspitz, D-82152 Martinsried, Germany.

*To whom correspondence should be addressed.

Fig. 63 In 1994, our work published on Solving the MOP gene was the basis for its correspondent profein's 3D determination.

MOP

The molybdenum enzymes are ubiquitous proteins in a variety of species, classified according to their molybdenum cofactor. The nitrogenases contain the iron-molybdenum cofactor (FeM-oco) and all other known enzymes contain the molybdopterin cofactor (Moco). The molybdenum enzymes catalyze redox reactions like xanthine dehydrogenase, sulphite oxidase, nitrate reductase, dimethylsulfoxide reductase and formate dehydrogenase. Additional cofactors may be present in the molybdenum enzymes, as for example flavin, b-type cytochrome and iron-sulphur centers. The molybdenum iron-sulphur protein (MOP) first described by Moura *et al*. in 1976 is an aldehyde oxidase from *Desulfovibrio gigas*. Romão *et al*, 1993 stated that the MOP contained two [2Fe-2S] centers as additional cofactors and has a molecular mass of approximately 100 kDa. (Thoenes *et al*, 1994) It's located in an operon containing a total of seven ORF's, namely a putative molybdopterin enzyme and a putative nitroreductase, as shown in Fig. 64 and 65.

Figure 64 – *Overview over the assembled and annotated genome. The red circle is signing the region where the MOP and surroundings are situated, obtained from artemis software.*

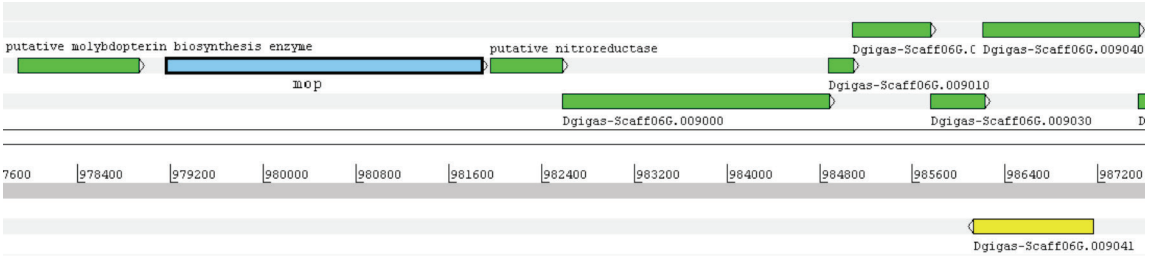


Figure 65 – *Region containing the MOP gene and surrounding area, obtained from artemis software*

Eur. J. Biochem. 220, 901–910 (1994)

Molecular cloning and sequence analysis of the gene of the molybdenum-containing aldehyde oxido-reductase of *Desulfovibrio gigas*

The deduced amino acid sequence shows similarity to xanthine dehydrogenase

Ulrich THOENES[1,2], Orfeu L. FLORES[1], Ana NEVES[1], Bart DEVREESE[3], Jozef J. VAN BEEUMEN[3], Robert HUBER[2], Maria J. ROMÃO[4], Jean LeGALL[6], José J. G. MOURA[5] and Claudina RODRIGUES-POUSADA[1]

[1] Instituto Gulbekian de Ciênca, Laboratório de Genética Molecular, Oeiras, Portugal
[2] Max-Planck-Institut für Biochemie, Martinsried, Germany
[3] Vakgroep Biochemie, Fysiologie en Microbiologie, Gent, Belgium
[4] Instituto de Tecnologia Quimica e Biológica, Oeiras and Instituto Superior Tecnico, Dep. Quimica, Lisboa, Portugal
[5] Departamento de Química, FCT, Universidade Nova de Lisboa, Monte da Caparica, Portugal
[6] Department of Biochemistry, University of Georgia, Athens, Georgia, USA

(Received December 6, 1993/January 14, 1994) – EJB 93 1808/2

In this report, we describe the isolation of a 4020-bp genomic *Pst*I fragment of *Desulfovibrio gigas* harboring the aldehyde oxido-reductase gene. The aldehyde oxido-reductase gene spans 2718 bp of genomic DNA and codes for a protein with 906 residues. The protein sequence shows an average 52% (±1.5%) similarity to xanthine dehydrogenase from different organisms. The codon usage of the aldehyde oxidoreductase is almost identical to a calculated codon usage of the *Desulfovibrio* bacteria.

The molybdenum enzymes are ubiquitous proteins in a variety of species e.g. bacteria, plants, animals and man. They are classified according to their molybdenum cofactor. The nitrogenases contain the iron-molybdenum cofactor (FeMoco) as described recently [1, 2]. All other known enzymes contain the molybdopterin cofactor (Moco). They catalyze redox reactions like xanthine dehydrogenase, sulphite oxidase, nitrate reductase, dimethylsulfoxide reductase and formate dehydrogenase. Additional cofactors may be present in the molybdenum enzymes; i.e. flavin, *b*-type cytochrome and iron-sulphur centers. The molybdenum iron-sulphur protein (MOP) first described by Moura et al. [3] is an aldehyde oxidase [4] from *Desulfovibrio gigas*. *D. gigas* is a sulphate-reducing, strictly anaerobic Gram-negative bacterium [5]. MOP contains two [2Fe-2S] centers as additional cofactors and has a molecular mass of approximately 100 kDa [6]. The first 34 N-terminal amino acids of MOP have been determined by amino acid sequence analysis [6]. A related protein has been isolated from *D. desulfuricans* ATCC 27774 (our unpublished results), but no sequence information is available. So far 25 proteins of the *Desulfovibrio* family are known by their primary structure. The genes of 22 proteins have been isolated.

In this work, we describe the isolation and sequence analysis of a gene coding for MOP. The deduced amino acid sequence of MOP is compared with other proteins containing the Moco cofactor. The codon usage of the MOP gene is compared to other genes of *D. gigas* and other species of *Desulfovibrio*.

Correspondence to U. Thoenes, Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany

Abbreviations. FeMoco, iron-molybdenum cofactor; Moco, molybdopterin cofactor; MOP, molybdenum-containing protein; DO(1–4), degenerated oligonucleotide.

Note. The novel amino acid and nucleotide data published here have been submitted to the EMBL sequence data bank.

MATERIALS AND METHODS

Bacterial strains and plasmids

D. gigas (NCIB 9332, DSM 1382) from Deutsche Sammlung von Mikroorganismen und Zellkulturen was grown anaerobically on Defined Multipurpose Medium as described by Widdel [7]. A 250-ml liquid culture was grown at 30°C for three days.

Isolated DNA fragments were cloned into the polylinker of pUC19 [8] and used for transformation of *Escherichia coli* strain JM109 [9]. Competent cells were prepared according to standard protocols [10].

Preparation and analysis of DNA

D. gigas genomic DNA was isolated as described elsewhere [11]. Plasmid DNA was prepared using the plasmid purification kit Qiagen (Diagen). For isolation and characterization of the gene of MOP *D. gigas* DNA was digested with *Pst*I and fragments of 3.5–4.5 kbp, positively hybridizing with an oligonucleotide derived from the N-terminus in a Southern-blot analysis were cut from the gel. DNA was extracted using the gel-extraction kit (Diagen). DNA was precipitated and resuspended in TE buffer (10 mM Tris, pH 8.0, 1 mM EDTA). About 60 ng DNA were ligated to 30 ng pUC19 which had been linearized with *Pst*I and dephosphorylated with alkaline phosphatase prior to ligation. Recombi-

Table 1. List of oligonucleotides and sequencing primers. I, inoside; R, A or G, Y, C or T.

Name	Position	Sequence
DO1	948–992	5′-ATGATYCARAARGTIATYACIGTIAAYGGIATYGARCARAAYCTG-3′
DO2	3458–3438	5′-GGAICCIGCICCIACIAGIGT-3′
DO3	1196–1176	5′-TTCIGGYTGICCIACICCTTC-3′
DO4	3130–3108	5′-TAIGTRCARTAICCICCICCIGG-3′
a	1327–1346	5′-GTGATTGGTTCCAGAAGCAT-3′
b	1691–1710	5′-GAACCGCATCACCGGCCTCA-3′
c	2119–2100	5′-CCGATGGACTTGGAGTGGAT-3′
d	2289–2308	5′-TACAACTATCAGCAGCAGCA-3′
e	2953–2934	5′-CAGGGCTTCGTGCGCCGTGC-3′
f	3084–3103	5′-AACCTCCTCAAGGCCTGTGA-3′
g	1821–1802	5′-TGGCTTCGGAGTCGGCGCAG-3′
h	1961–1980	5′-CAAGGGCGAAGACACCGGCC-3′
i	2641–2622	5′-TTGTAGCGCAGTTCCAGCGG-3′
k	2622–2641	5′-CCGCTGGAACTGCGCTACAA-3′-3′
l	3401–3421	5′-CCTGTCCGAGGACTTCGAGGA-3′
m	3421–3401	5′-TCCTCGAAGTCCTCGGACAGG-3′-3′
n	3671–3652	5′-AGGTTAGGCCTTCAAGGCTT-3′
o	986–967	5′-CTGCTCGATGCCGTTGACGG-3′
p	275–294	5′-CCCATTCTCAAGCGCAAGTT
Reverse primer	5′-side of pUC19 polylinker	5′-AACAGCTATGACCATG-3′
Universal primer	3′-side of pUC19 polylinker	5′-GTAAAACGACGGCCAGT-3′

nant plasmids were used to transform JM109 and plated on Luria-Bertani ampicillin plates containing 1 mM isopropyl β-D-galacto-pyranoside and 150 μg/ml 5-bromo-4-chloro-3-indolyl β-D-galacto-pyranoside (X-gal). Positive clones were selected and grown in overnight cultures.

Plasmids were isolated as described above, digested with *Pst*I, and electrophoresed on 0.8% agarose gels. For further analysis, the resulting plasmid pMOP1, positively hybridizing in Southern-blot analysis, was digested with *Eco*RI and with *Eco*RI/*Sac*I. The fragments obtained were subcloned into pUC19, resulting in pMOP2 to pMOP6. Fragments were also cloned in M13mp18/mp19.

The DNA sequences of the oligonucleotides (Table 1) were deduced from the peptide sequences of the N-terminus of MOP (DO1) and proteolytic fragments (DO2–DO4) or the DNA sequence information as it became available. Oligonucleotides and sequencing primers were synthesized on a DNA synthesizer from Applied Biosystems or Pharmacia LKB.

Labeling of oligonucleotides was in a 50-μl reaction volume containing 25 μCi [γ-^{32}P]dATP (1 mCi/ml; Amersham), 180 ng oligonucleotide and 20 U polynucleotide kinase (Pharmacia) for 1 h at 37°C. DNA fragments were labeled with 50 μCi [α-^{32}P]dCTP (1 mCi/ml; Amersham) using the mega-prime kit supplied by Amersham.

For Southern-blot analysis, chromosomal as well as plasmid DNA was digested with appropriate restriction enzymes and electrophoresed on an 0.8% agarose gels in TAE buffer (40 mM Tris/acetate, 2 mM EDTA, pH 8.5). After electrophoresis, DNA was transferred onto a Nylon Hybond N$^+$ membranes (Amersham) by capillary transfer using 0.4 M NaOH according to standard protocols [12]. Membranes were prehybridized using 6×NaCl/Cit (0.9 M NaCl, 90 mM sodium citrate, pH 7.0), 0.5% SDS, 5×Denhardt's solution (100×Denhardt solution is 10 g Ficoll 400, 10 g poly(vinylpyrrolidone), 10 g bovine serum albumin in 500 ml H_2O) and 100 μg/ml sonicated salmon sperm DNA for 4 h at 42°C.

Hybridization was performed in 20 ml 6×NaCl/Cit, 1× Denhardt's solution, 100 μg/ml sonicated salmon sperm DNA and 50 μg/ml $Na_2P_4O_7$ for 16 h with the addition of the ^{32}P-labeled probe. For the oligonucleotides DO1 and DO2 as probes, the hybridization temperature was 56°C, for the labeled 981-bp *Eco*RI fragment it was 65°C.

Membranes were either washed in 6× NaCl/Cit at 25, 30 and 40°C (DO1), or at 25, 28 and 35°C (DO2) each for 10 min. Membranes were washed for 10 min in 2×NaCl/Cit, 0.1% SDS at approx. 50°C when DNA fragments were used as probes. Plasmids were sequenced using the supplied primers for the M13/pUC-system (Pharmacia and Boehringer) and also making use of primers which were derived from parts of the sequence after sequencing (Table 1). The 984-bp and 1400-bp *Eco*RI fragments were also cloned in the replicative form of M13mp19. Double-stranded as well as single-stranded DNA was sequenced with the dideoxy-chain-termination procedure [13] using the T7-sequencing kit (Pharmacia). To improve the gel reading of the C-terminal fragments which are rich in G/C, the terminal deoxynucleotidyl transferase technique without dGTP analogues was used. Sequencing experiments using analogues of dGTP (dITP and 7-deazaGTP were used in parallel) were performed in addition whenever a compression was suspected.

Computer methods

DNA and protein analysis were performed either on a DEC-VAX computer using the UWGCG-system (Genetics Computer Group) or on a MS-DOS PC using the DNASIS-program (Pharmacia LKB).

Peptide sequencing

Internal sequences of MOP were obtained using two different digests. 8 nmol native protein, purified as described [14], were used for the digestion with *Lysobacter enzymo-*

Table 2. List of Lys-C protease proteolytic and acid hydrolysis peptides of MOP. Molecular mass is only shown in those cases where it was determined by mass spectroscopy. Numbers 1, 2, 12, 13, 21, 22 are overlapping peptides.

Number	Location	N-terminal sequences found by peptide sequencing; (These do not always match the corresponding peptide sequence deduced from the DNA sequence)	Molecular mass (experimental by MS)	Molecular mass (calculated from DNA deduced peptides, range in brackets)
			Da	
N-term	1– 34	MIQKVITVNGIEQNLFVDAEALLSDVL(R)Q(Q)– –LT		
1	54– 68	GKVVRA?VTKMKKVA	1760.62	1760.00 (54– 69)
2	66– 89	GVA(D)GAQITTIEGVGQPENL?P(L)?	2671-46	2670.42 (65– 90)
3	113–130	GLLDTNADPS?E(E)V(V)?DF(FDQ)??	2360.27	2361.14 (133–132)
4	144–158	(PLVDA)VM(Y)A(A)AV(I)NG	1583.04	1583.87 (144–159)
5	160–167	KPETDLEFK	1105.38	1106.25 (160–168)
6	169–178	MPADG?I(E)GS?(NLN)	1214.88	1216.61 (169–179)
7	180–188	YPRPTAVAK	1001.15	1001.57 (180–188)
8	189–200	V(?)G(T)L?YGADLG	1422.60	1421.75 (189–202)
9	203–208	MPAG–LI–FP		
10	217–222	(V)SAANI		
11	224–240	GIDTSEALTM(P)GVVSSI?V?	3216.32	3216.63 (224–243)
12	245–257	VKGKNRITGLITF		
13	249–259	(NA)F(GG)LI(T)F(P)P–(Y)P		
14	267–274	RPILßDEK		
15	275–284	VFQYG???AL		
16	289–308	?EANARAAAEKVKVPLEELP		
17	310–329	YMSGPAA(W)?E(L)AI(T)I?PGT(P)		
18	358–370	FYVGRQPHMPIEP	1686.03	1684.81 (358–371)
19	389–403	(S)I?VYLNLYMIA(PP)V	3795.54	3792.98 (389–424)
20	425–443	FSPTSEALVAV(AA)M?TG(R)?P	3912.35	3909.97 (425–459)
21	559–568	PLELRYKNAY	1691.10	1790.89 (559–572)
22	566–587	NAYRPGDTNPTGQ(E)P(F)VFFLPD	3386.73	3384.65 (566–595)
23	596–602	YQAALEK		
24	721–732	PGGGY(C)TYDG(LT)K	1105.38	1106.25 (721–732)
25	736–748	KPTKIGN?TA?G	1887.70	1886.90 (736–753)
26	829–841	(R)ATLVGAGF(PF)IPNI(Y)GL		
27	842–845	QIPD		
28	849–857	IVYVNHPRP	1208.85	1208.64 (849–858)
29	859–872	(G)PFGASGVGTLPLT	2416.90	2417.80 (859–882)

genes lysyl-C protease (Wako; E/S 1 : 100). The peptide mixture was separated on a RoSil TMS 3-µm column (Bio-Rad) using a DuPont Instruments separation system consisting of a 870 chromatographic pump, a 8800 gradient controller and an ultraviolet spectrophotometer. Precipitation was observed after this digestion. The precipitate was further digested by partial acid hydrolysis with 2% formic acid at 108 °C for 2 h. We also performed partial acid hydrolysis on the total protein under the same conditions as for the precipitate. Peptide mixtures in all cases were chromatographed on an PTC C_{18} column using a 140A solvent delivery system and 1000S diode array detector (all Applied Biosystems).

Sequence analysis was performed on either a 475A or a 477A pulsed-liquid protein sequencer with on-line analysis of the phenylthiohydantoin–amino-acids on a 120A analyser (all Applied Biosystems):

Mass spectra

The electrospray mass spectra were obtained on 100 pmol sample dissolved in 10 µl 50% acetonitrile/1% formic acid in water. Samples were introduced into the mass spectrometer using a flow rate of 6 µl/min, pumped by a 140A solvent delivery system (Applied Biosystems). The mass spectrometer was a VG BIO-Q triple quadrupole instrument (Fisons Instruments). Spectra were obtained scanning *m/z* values over 600–1500 during 9 s using only the first quadrupole as the mass analyser. Scans were accumulated for 3 min. Calibration was performed using horse myoglobin (Sigma).

RESULTS AND DISCUSSION

Isolation and analysis of the gene encoding the MOP of *D. gigas*

The N-terminal sequences determined for all 29 peptides obtained from the MOP digest using Lys-C protease are summarized in Table 2, together with 34 amino acids of the N-terminus of MOP [6]. The sequence data were used to construct oligonucleotides. These degenerated oligonucleotides were designed to be complementary to the combinations of mRNA that could encode the protein. We have therefore taken into account the codon usage of *D. gigas* as deduced from the genes for desulforedoxin [15], flavodoxin [16] and [NiFe]hydrogenase [17].

Starting with the methionine of the N-terminus, a 45-bp-long degenerated oligonucleotide was synthesized (DO1). Three additional degenerated oligonucleotides (DO2–DO4) were synthesized based on the information obtained from the

sequence data of the 'Lys-C peptides'. Additional oligonucleotides designated a–p (Table 2, Fig. 2) were synthesized for complete sequence analysis.

As an initial step to clone the MOP gene from *D. gigas*, we have examined the ability of the probe DO1, (see Table 2) to hybridize the genomic DNA. DNA was thus digested with *Sac*I, *Pst*I, *Bam*HI, *Hin*dIII and *Bam*HI, and the fragments were separated by electrophoresis on an 0.8% agarose gel. The DNA fragments were transfered to nylon membranes and hybridized with labeled DO1 as described in Materials and Methods. The hybridization pattern is shown in Fig. 1a, revealing one single band in each lane. The expected size of the gene of MOP is about 2500-2600 bp, corresponding to the band 1 in lane 5 (*Eco*RI digest).

As we did not expect that the entire gene was located on this fragment, we also isolated longer fragments corresponding to the band at approximately 4 kb in lane 2 (*Pst*I digest). The DNA fragments from the region 3.5–4.5 kb were cut from the preparative gel and subcloned into pUC19. 40 white colonies were obtained and 20 randomly selected which were further analyzed by digestion with *Pst*I. After Southern blotting and hybridization using labeled DO1, two of them labeled, which were identical. The clone was designated pMOP1. Plasmid pMOP1 was further digested with *Eco*RI, *Eco*RI–*Hin*dIII, *Eco*RI–*Pst*I, *Eco*RI–*Sac*I, *Hin*dIII–*Sac*I, *Pst*I–*Sac*I and *Sac*I. The partial restriction map of pMOP1 deduced from the result is shown in Fig. 2. To localize the start codon of the gene on one of the restriction fragments, the gel was Southern blotted. The resulting membrane was hybridized with labeled oligonucleotide DO1 as probe. The start codon was identified on the *Eco*RI–*Pst*I fragment (approximately 1440 bp). The coding direction of the gene in pMOP1 was identified by hybridizing a similar Southern blot with labeled DO2 as probe. DO2 is localized on the 1166-bp *Pst*I–*Sac* fragment. In this way, the coding direction of the MOP gene on the cloned *Pst*I fragment was determined as shown in Fig. 2. Using the results from the described digests five different fragments of pMOP1 were subcloned generating pMOP2–pMOP6.

pMOP2:	approx. 1440 bp	(*Eco*RI–*Pst*I)
pMOP3:	981 bp	(*Eco*RI–*Eco*RI)
pMOP4:	1597 bp	(*Eco*RI–*Pst*I)
pMOP5:	431 bp	(*Eco*RI–*Sac*I)
pMOP6:	1166 bp	(*Sac*I–*Pst*I)

pMOP1–pMOP6 were used to determine the complete DNA sequence of the MOP gene including its flanking regions. The total length of the *Pst*I fragment was determined to be approximately 4020 bp harboring the 2721-bp long coding region of MOP. The sequence is shown in Fig. 3, together with the resulting amino acid sequence of MOP and the location of the peptides. Sequencing was started with pMOP1 using the reverse primer and the universal primer (see Table 1). Within the resulting sequence information, none of the given peptides could be localized. The sequencing was continued with clone pMOP2 using the same primers as well as oligonucleotide DO1. Here the positions of peptides 1 and 2 could be found. Using oligonucleotide DO3 as sequencing primer, the N-terminus was confirmed and precisely located. The site of the start codon was found to be 502 bp upstream of the *Eco*RI site of pMOP2. During further improved sequencing reactions, the peptides 3–5 could also be discovered in the part of the gene cloned in pMOP2.

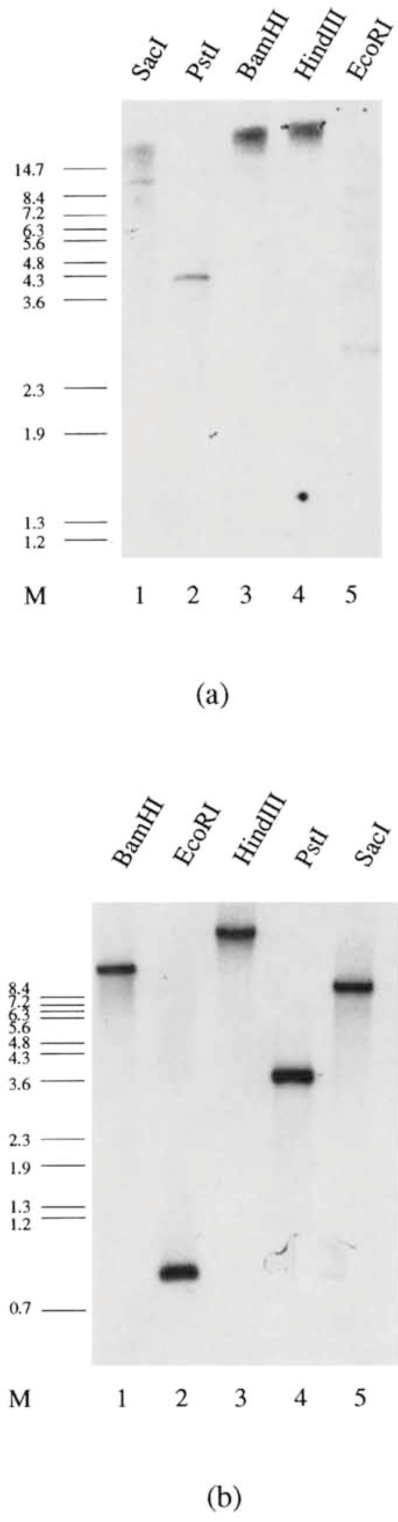


Fig. 1. Autoradiogram of two Southern blots. Genomic DNA of *D. gigas* was digested with *Sac*I, *Pst*I, *Bam*HI, *Hin*dIII, and *Eco*RI. Hybridization was with labeled oligonucleotide DO1 (a) and with the labeled 981-bp *Eco*RI fragment isolated from pMOP3 (b). M; marker in kb; *Bst*II digest of λ DNA

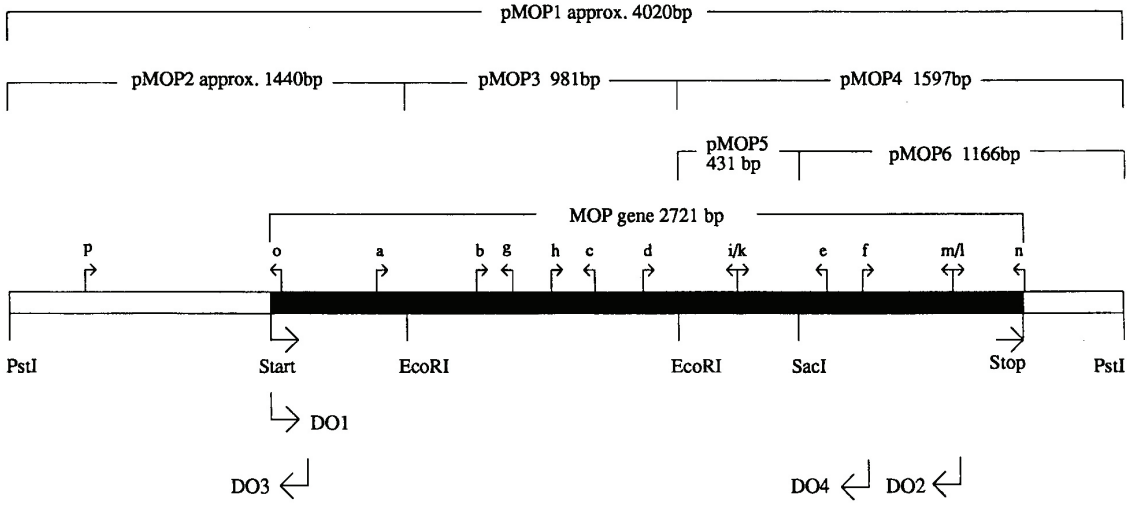


Fig. 2. Partial restriction map and organization of cloned fragments. Subcloned fragments are shown with the names of the corresponding plasmids and their length. The synthesized primers (all 20 nucleotides long) and their orientation are shown above the sequence box by little arrows and small letters above them. The degenerated oligonucleotides DO1−DO4 are shown below the sequence with arrows indicating their 5′-3′-orientation.

Sequence analysis was continued with pMOP3, again using the universal primer and the reverse primer. About two thirds of the sequence part cloned into pMOP3 could be determined using those two primers. The missing part between them could be sequenced on both strands using primer b and primer c.

In this 981-bp-long part of the sequence, the positions of the majority of the peptides, namely peptides 6−20, were found (see Table 2).

Using pMOP4 and pMOP5, the sequence of the smallest cloned part of the gene the 431-bp *Eco*RI−*Sac*I fragment was determined using the universal primer, the reverse primer, primers i and k. In this fragment, three of the given peptides namely 21−23 were found.

Finally, the 1166-bp *Sac*I−*Pst*I fragment in pMOP6 was sequenced using the reverse, the universal primer, the primers f, l, m, n and the degenerated oligonucleotide DO3. In this sequence part, the positions peptides 24−29 and the stop codon of the MOP protein were found.

In order to confirm the two *Eco*RI sites and the *Sac*I site, specified regions of pMOP1 were sequenced using the primers a, d and e.

As outlined in Table 2, some amino acid sequences obtained from the peptides did not exactly match the DNA sequence. In those cases the molecular masses of the peptides were determined by mass spectroscopy. These corresponded very closely to the values deduced from the gene sequence, providing good evidence for its accuracy. All the given peptides were found in the deduced amino acid sequence of MOP. They are unequally distributed over the whole sequence. Most of them are clustered in two regions extending from P144−P329, harboring peptides 4−17, and from H829−T872 containing peptides 26−29. This may indicate that Lys in these regions are more exposed and therefore more accessible for the Lys-C protease.

The molecular mass of MOP was also determined by electrospray ionisation mass spectroscopy to be 97210.7 Da. This is only 76.18 Da smaller than the value of 97286.88 Da deduced from the gene, a difference of less than 0.1% in accord with the expected accuracy of the mass spectrometric method [18]. These values are also in good agreement with electrophoresis and sedimentation data [6].

Codon usage

The DNA-sequence information was used to examine the codon usage of the MOP gene. It was compared to the codon usages of the three known genes of *D. gigas* and of all known genes of the *Desulfovibrio* bacteria (22 genes). Considering twofold degeneracy, MOP *D. gigas* and *Desulfovibrio* show identical codon usage, with the exception of Gln and Asp. At higher degeneracy the two most often used codons are identical with the exception of Arg, Ala and Gly where only the first most used codon is the same.

Overall, the frequency of the two most often used codons is in a range 50−70%. Major differences in the codon frequencies are only present in the remaining codons which are not used often. It should be noted that the relative frequency of the triplets AGA(R), ATA(I), ACA(T), AGT(S), TTG TTA CTA(L) is low (below 10%) and of GAA(E), GAT(D), GCC(A), AAG(K), AAC(N), TGC(C), TTC(F); CAA(Q) is high (above 50%) compared to the other triplets coding for the same amino acid.

Genomic organization

To confirm the Southern-blot pattern of DO1, additional Southern-blot analysis was performed. We now used the 981-bp *Eco*RI-fragment (Fig. 2) from pMOP3 which was labeled using Klenow fragment of DNA polymerase I. The results are shown in Fig. 1b.

As the *Eco*RI fragment is part of all restriction fragments containing the MOP gene, the hybridization pattern is similar as that of DO1 (Fig. 1a), with the exception of *Eco*RI itself. Here, we expect to have hybridization with the intra-gene fragment, while DO1 hybridizes with a larger 5′-fragment. Due to the longer and not degenerated DNA fragment used as probe, the bands are much more intense.

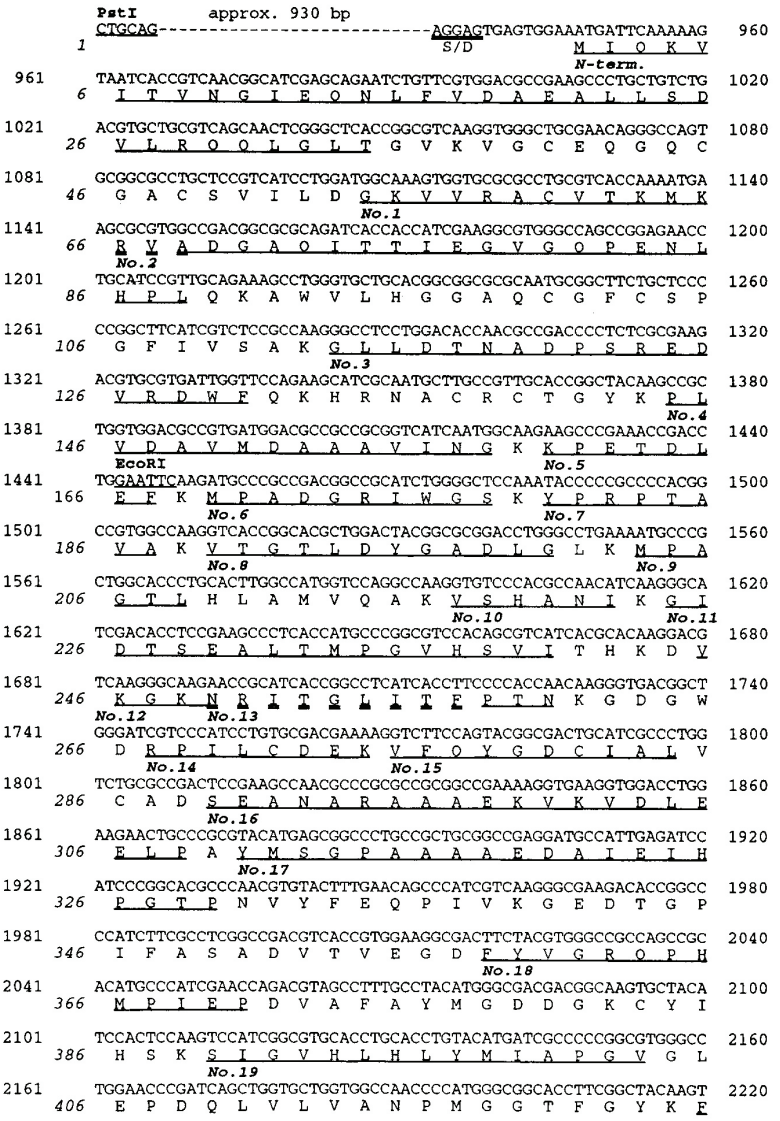


Fig. 3. DNA and deduced amino acid sequence of the *Pst*I fragment harboring the complete MOP gene. The 'LysC' and acid hydrolysis peptides (see text) are underlined and numbered in italic. Restriction sites are shown above the DNA sequence.

Our results clearly indicate that the aldehyde oxidoreductase is present as a single copy gene on the genome of *D. Gigas*.

Further sequence analysis of the region upstream of the coding region of MOP (not shown) did not result in any information about promotor-like regions. This data are strong indications for a polycistronic organization of the MOP gene.

Sequence alignment with xanthine dehydrogenase

The deduced amino acid sequence of MOP was used in a database search. Four different xanthine dehydrogenases from rat, mouse, *Drosophila melanogaster* and *Caliphora vicina* were found to give the highest similarity scores. The average similarity between MOP and the xanthine dehydrogenases was found to be 52% (25.6% identity). In xanthine dehydrogenases [19], the binding of molybdenum to the protein is mediated by the pterin cofactor. A NAD/FAD-binding domain is suggested to be located N-terminal to the putative Mo-pterin-binding domain. In addition, two different [2Fe-2S] domains are located in the N-terminal segment. The four amino acid sequences of xanthine dehydrogenase were aligned with MOP.

The result can be seen in Fig. 4. In two regions, 16–179 and 627–1364 (general numbers) the sequences show particularly high similarity. The segment 180–626 encompasses the putative flavin-binding domain in xanthine dehydrogenase. In MOP for which no flavin cofactor has been demonstrated, this region is absent except for four short peptide regions of only 12–18 amino acids which are positioned rather arbitrarily in Fig. 4. Due to the lack of sequence information, it was not possible to compare MOP to the aldehyde oxidase from rabbit liver [20] containing molybdenum, nonheme iron and FAD. Furthermore, a comparison between

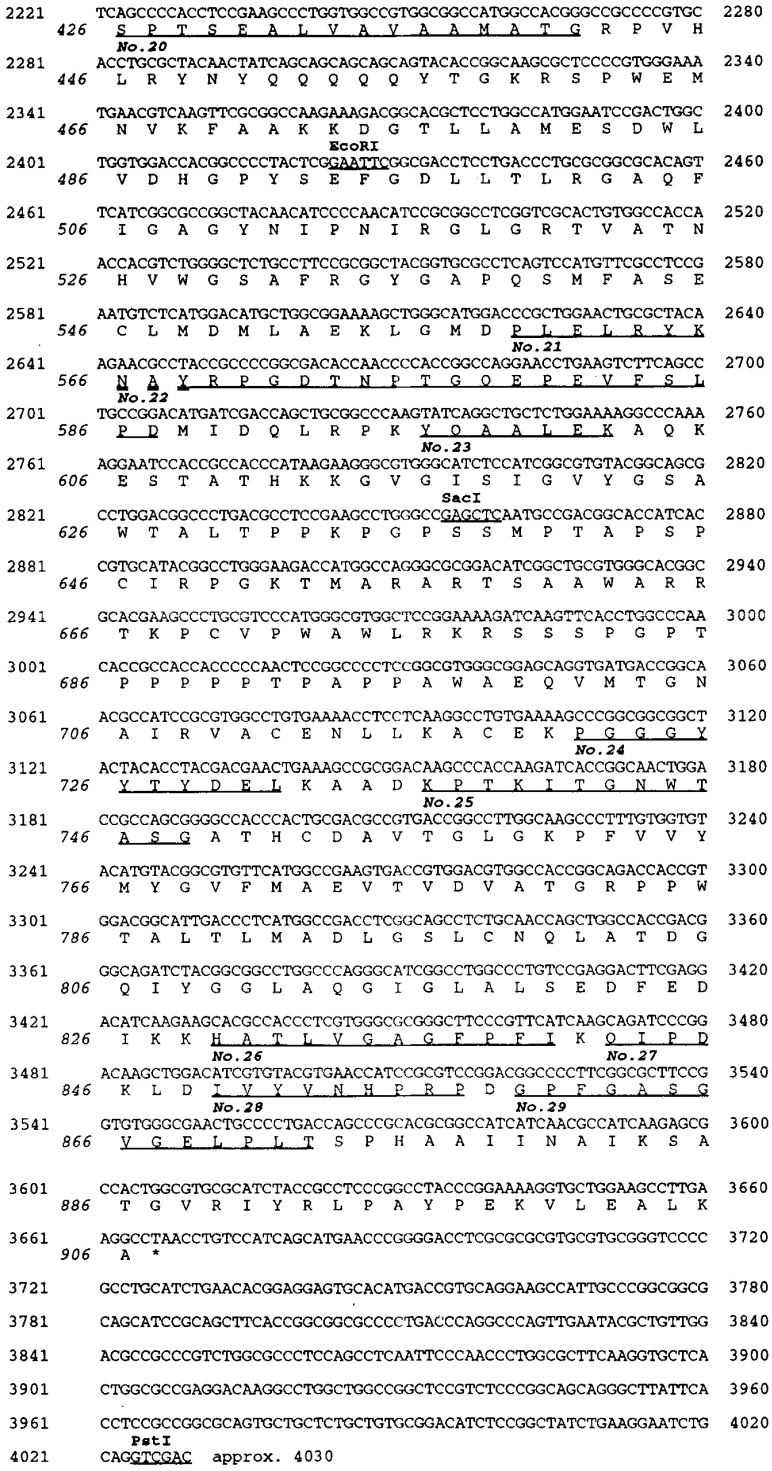


Fig. 3. (Continued).

MOP and a putative FAD-free aldehyde oxidase containing tungsten an Fe/S-centers from *T. litoralis* [21] could not be carried out because of the missing sequence information.

In the N-terminal region of the alignment, eight well conserved cysteines were found. The first four cysteines (58, 63, 66 and 88; general numbers) carry the signature of plant ferredoxins *(Aphanothece sacrum* and *Spirulina platensis* [22, 23] where the four cysteines ligating the iron atoms are at the positions 41, 46, 49 and 79 [23]. They might, therefore, be considered to bind the first [2Fe-2S] center. The second

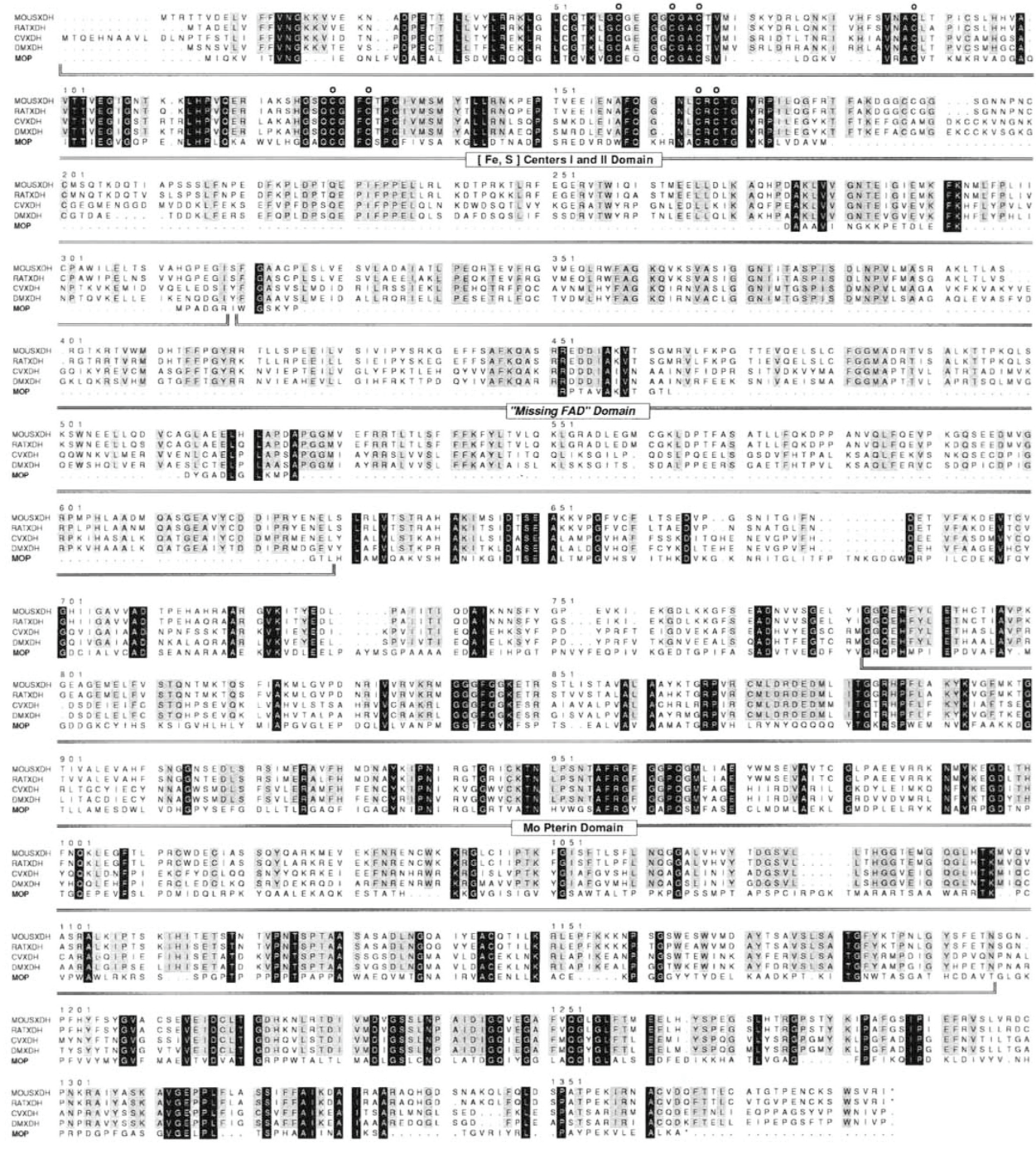


Fig. 4. Alignment of MOP and xanthine dehydrogenases from mouse (MOUSXDH), rat (RATXDH), *C. vicina* (CVXDH), *D. melanogaster* (DMXDH). Conserved amino acids are marked by black background. The cysteins binding the [2Fe-2S] centers are marked by "○".

four well conserved cysteines (129, 132, 166, 168; general numbers) might bind to the second [2Fe-2S] center, but do not show similarity to other structurally defined iron-sulfur proteins. In the C-terminal region of the alignment, the putative molybdo-pterin-binding domain [19] of the xanthine dehydrogenases (747−1256; general numbers) aligns well with the MOP sequence, suggesting similar structural cofactor-binding properties. Significant similarities to other prokaryotic molybdopterin-containing proteins [19] were not found.

Its similarity to xanthine dehydrogenase suggests that MOP might have a similar enzymic activity to that already reported by some of us [4]. Xanthine dehydrogenase cata-

lyzes the oxidation of xanthine to hypoxanthine and subsequently to uric acid with simultaneous reduction of NAD to NADH. This so called D-type activity (NAD reductase activity) in the xanthine dehydrogenase may be converted irreversibly or reversibly to O-type activity (O_2 reductase activity) in xanthine oxidase [24]. Irreversible conversion can be achieved by treatment of xanthine dehydrogenase with trypsin, resulting in decreased NAD-binding activity of xanthine dehydrogenase [25]. The lack of NAD binding in MOP would similarly suggest an O-type activity for this enzyme. However *D. gigas* is a strictly anaerobic bacterium which excludes the O-type reaction. The physiological electron acceptor of MOP is not known and may be NAD or menaquinone which is present in *D. gigas* [26] and assumed to be the acceptor for the malate dehydrogenase of the *Desulfobacter* bacteria [27]. It may be that a separate cofactor-binding protein takes the role of the flavoprotein domain of xanthine dehydrogenase in *D. gigas*.

However, a flavo-hemoprotein has been purified from *D. gigas* that reduces O_2 to water and this activity is linked to NADH oxidation [28]. It should be noted that, although classified as a strict anaerobe, *D. gigas* has the capability of synthesizing nucleotide triphosphate from the degradation of polyglucose in the presence of oxygen [29]. Based on the observation that glyceraldehyde is a substrate for MOP, it has been recently proposed that its physiological role is linked to the degradation of polyglucose by *D. gigas* [30]. The protein is capable of reducing flavodoxin and cytochrome c_3; H_2 can be produced from aldehydes in the presence of hydrogenase, an essential component in this complex electron-transfer chain. Thus, any eventual link between MOP and pyridine nucleotides remains to be demonstrated.

We acknowledge the support of the following institutions: *Kernforschungszentrum Karlsruhe, Stabsabteilung internationale Beziehungen (Projektnummer: X1844), Junta Nacional de Investigação Científica e Tecnológica (Portugal), and Fundação Calouste Gulbenkian*. O. L. Flores received a fellowship from J. N. I. C. T., Jozef J. Van Beeumen is indebted to the Belgium Fund for Joint Basic Research (contract 3.0018.91), and to the Belgium Fund for Medical Scientific Research (contract 39.0038.91). We thank M. J. Feio and I. Moura for experimental contributions and M. Archer for helping with the representation of the sequence alignment. Prof. Dr F. Widdel and Dipl.-Biol. P. Tormay are gratefully acknowledged for advise in growing *D. gigas*. Special thanks to Prof. Dr B. Brenig and Dipl.-Biol. J. Stiebler for experimental advise.

REFERENCES

1. Kim, J. & Rees, D. C. (1992) Structural Models for the Metal Centers in the Nitrogenase Molybdenum-Iron Protein, *Science 257*, 1677–1682.
2. Kim, J. & Rees, D. C. (1992) Crystallographic structure and funtional implications of the nitrogenase molybdenum-iron protein from *Azetobacter vinelandii*, *Nature 360*, 553–560.
3. Moura, J. J. G., Xavier, A. V., Burschi, M., Le Gall, J., Hall, D. O. & Cammack, R. (1976) A Molybdenum-Containing Iron-Sulfur Protein from *Desulfovibrio gigas*, *Biochem. Biophys. Res. Commun. 72*, 782–789.
4. Turner, N., Barata, B., Bray, R. C., Deistung, J. & Le Gall, J. (1987) The molybdenum iron-sulphur protein from *Desulfovibrio gigas* as a form of aldehyde oxidase, *Biochem. J. 243*, 755–761.
5. Le Gall, J. (1963) A new Species of *Desulfovibrio*, *J. Bacteriol. 86*, 1120.
6. Romao, M. J., Barata, B. A. S., Archer, M., Lobeck, K., Moura, I., Carrondo, M. A., LeGall, J., Lottspeich, F., Huber, R. & Moura, J. J. G. (1993) Subunit composition, crystallization and preliminary crystallographic studies of the *Desulfovibrio gigas* aldehyde oxidoreductase containing molybdenum and [2Fe-2S] centers, *Eur. J. Biochem. 215*, 729–732.
7. Widdel, F. & Bak, F. (1992) in *The prokaryotes* (Barlow, A., ed.) vol. 3, pp. 3351–337, Springer-Verlag, New York.
8. Messing, J. (1983) New M13 vectors for cloning, *Methods Enzymol. 101*, 20–83.
9. Yanisch-Perron, C., Vieira, J. & Messing, J. (1985) Improved M13 phage cloning vectors and host strains: Nucleotide sequences of the M13mp18 and pUC19 vectors, *Gene (Amst.) 33*, 103–119.
10. Hanahan, D. (1983) Studies on Transformation of *Escherichia coli* with plasmids, *J. Mol. Biol. 166*, 557–580.
11. Ausubel, F. M., Brent R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1990) *Current protocols in molecular biology*, 2.4.1.–2.4.2.
12. Southern, E. (1975) Detection of Specific Sequences Among DNA-Fragments Separated by Gel Electrophoresis, *J. Mol. Biol. 98*, 503–517.
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) DNA-Sequencing with Chain-Terminating Inhibitors, *Proc. Natl Acad. Sci. USA 74*, 5463–5467.
14. Moura, J. J. G., Xavier, A. V., Cammack, R., Hall, D. O., Burschi, M. & Le Gall, J. (1978) Oxidation-Reduction Studies of the Mo(2Fe-2S) Protein from *Desulfovibrio gigas*, *Biochem. J. 173*, 419–425.
15. Brumlik, M. J., Leroy, G., Bruschi, M. & Voordouw, G. (1990) The Nucleotide Sequence of the *Desulfovibrio gigas* Desulforedoxin Gene Indicates that the *Desulfovibrio vulgaris rbo* Gene Originated from Gene fusion Event, *J. Bacteriol. 172*, 7289–7292.
16. Helms, L. R. & Swenson, P. S. (1992) The primary structures of the flavodoxins from two strains of *Desulfovibrio gigas*. Cloning and Nucleotide sequence of the structural genes. *Biochim. Biophys. Acta 1131*, 325–328.
17. Li, C., Peck, H. D., LeGall, J. & Przybyla, A. E. (1987) Cloning, Characterization, and Sequencing of the Genes Encoding the Large and Small Subunits of the Periplasmatic [NiFe]hydrogenase of *Desulfovibrio gigas*, *DNA 6*, 539–551.
18. Smith, R. D., Loo, J. A., Edmonds, C. G., Barinaga, C. J. & Udseth, H. R. (1990) New Developments in Biochemical Mass Spectroscopy: Electrospray Ionization, *Anal. Chem. 62*, 882–899.
19. Wootton, J. C., Nicolson, R. E., Cock, J. M., Walters, D. E., Burke, J. F., Doyle, W. A. & Bray, C. (1991) Enzymes depending on the pterin molybdenum cofactor: sequence families, spectroscopic properties of molybdenum and possible cofactor-binding domains, *Biochim. Biophys. Acta 1057*, 157–185.
20. Branzoli, U. & Massey, V. (1974) Preparation of Aldehyde Oxidase in Its Native and Deflavo Forms, *J. Biol. Chem. 249*, 4339–4345.
21. Mukund, S. & Adams, M. W. W. (1993) Characterization of a Novel Tungsten-containing Formaldehyde Ferredoxin Oxidoreductase from the Hyperthermophilic Archaeon, *Thermococcus litoralis*, *J. Biol. Chem. 268*, 13592–13600.
22. Amaya, Y., Yamazaki, K., Sato, M., Noda, K., Nishino, T. & Nishino, T. (1990) Proteolytic Conversion of Xanthine Dehydrogenase from NAD-dependent Type to O_2-dependent Type, *J. Biol. Chem. 265*, 14170–14175.
23. Tsutsui, T., Tsukihara, T., Fukuyama, K., Katsube, Y., Hase, T., Matsubara, H., Nishikawa, Y. & Tanaka, N. (1983) Main Chain Fold of a [2Fe-2S]Ferredoxin I from *Aphanothece sacrum* at 2.5 Å Resolution, *J. Biochem. 94*, 299–302.
24. Terao, M., Cazzaniga, G., Ghezzi, P., Bianchi, M., Falciani, F., Perani, P. & Garattini, E. (1992) Molecular cloning of a cDNA coding for mouse liver xanthine dehydrogenase, *Biochem. J. 283*, 863–870.
25. Amaya, Y., Yamazaki, K., Sato, M., Noda, K., Nishino, T. & Nishino, T. (1990) Proteolytic Conversion of Xanthine Dehydrogenase from NAD-dependent Type to O_2-dependent Type *J. Biol. Chem. 265*, 14170–14175.

26. Postgate, J. R. (1979) *The sulphate-reducing bacteria*, pp. 60–69, Cambridge University Press.
27. Widdel, F. & Hansen, T. A. (1992) in *The prokaryotes* (Barlow, A., ed.) vol. 1, pp. 583–627, Springer Verlag, New York.
28. Chen, L., Lin, M.-Y., LeGall, J., Fareleira, P., Santos, M. H. & Xavier, A. V. (1993) Rubredoxin Oxidase, a new Flavo-Hemo-Protein, is the Site of Oxygen Reduction to Water by the "Strict Anaerobe" *Desulvovibrio gigas*, *Biochem. Biophys. Res. Commun. 193*, 100–105.
29. Santos, M. H., Fareleira, P., Xavier, A. V., Chen, L., Liu, M.-Y. & LeGall, J. (1993) *Biochem. Biophys. Res. Commun*, in the press.
30. Barata, B., LeGall, J. & Moura, J. J. G. (1993) Aldeyde Oxido-Reductase Activity in *Desulfovibrio gigas*, *Biochemistry 32*, 11559–11568.

IV.3 - Other metal protein's coding genes found in the genome

Following our studies with the MOP, and due to the importance of this metal in *D.gigas*, we decided to look for all the genome encoded proteins containing molybdenum and we found a total of 21, equally distributed on the genome as shown in the next figures. We also show the same search results from Ion protein and Tungsten proteins.

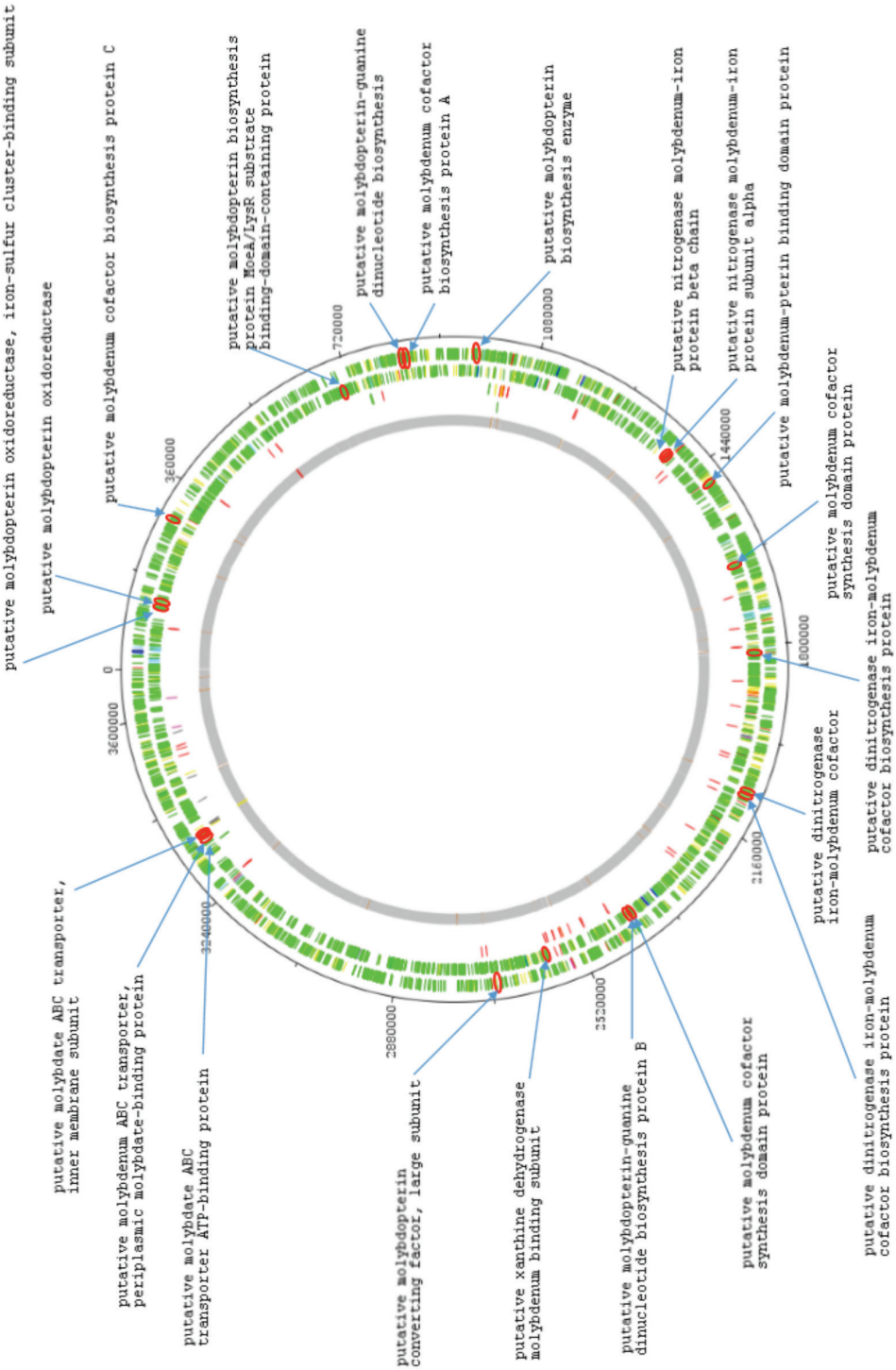


Figure 66 - *Localization of the genes coding for molybdenum containing enzymes on the annotated genome, obtained from artemis software.*

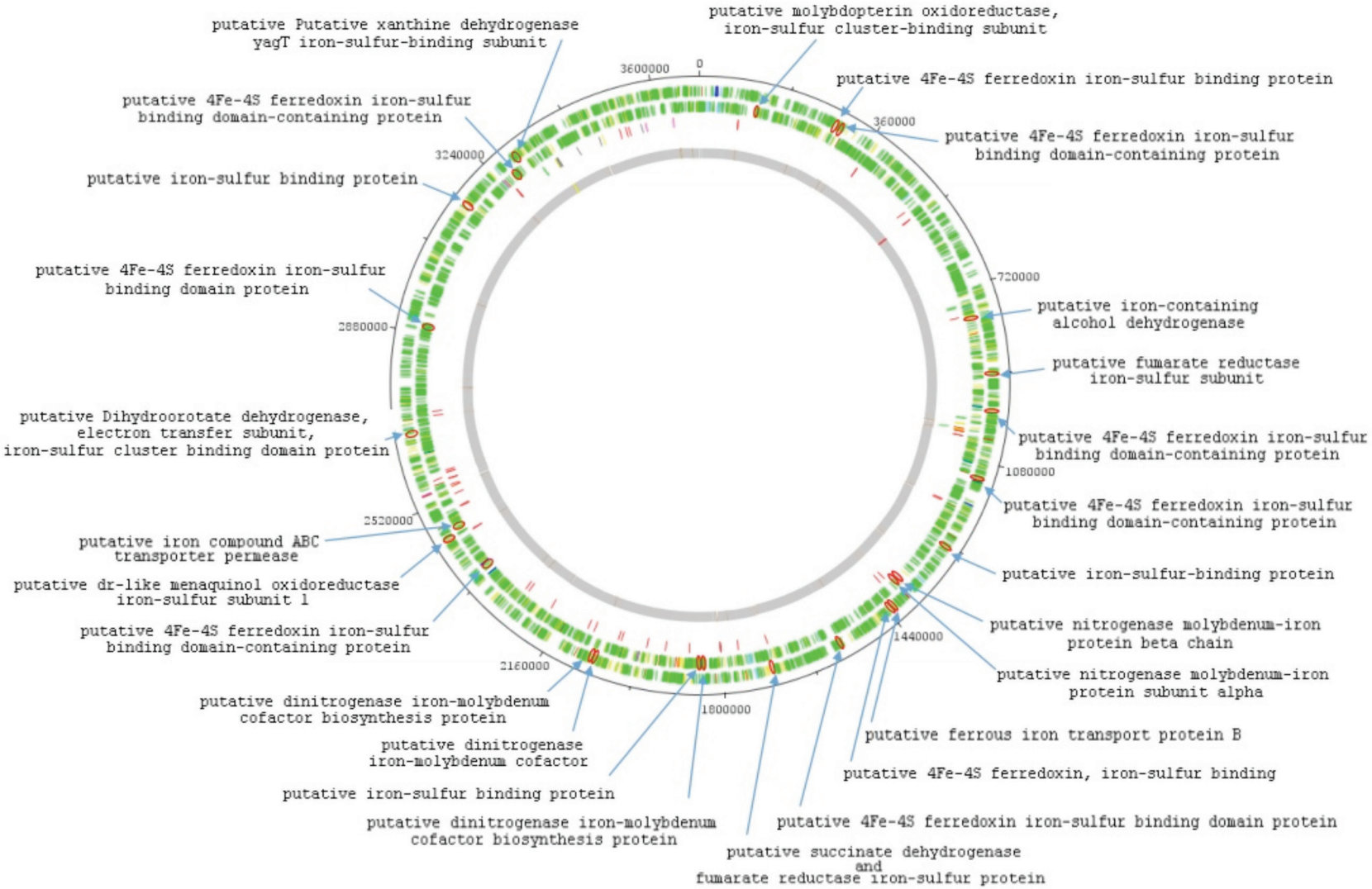


Figure 67 - *Localization of the genes coding for iron containing enzymes on the annotated genome, obtained from* artemis software.

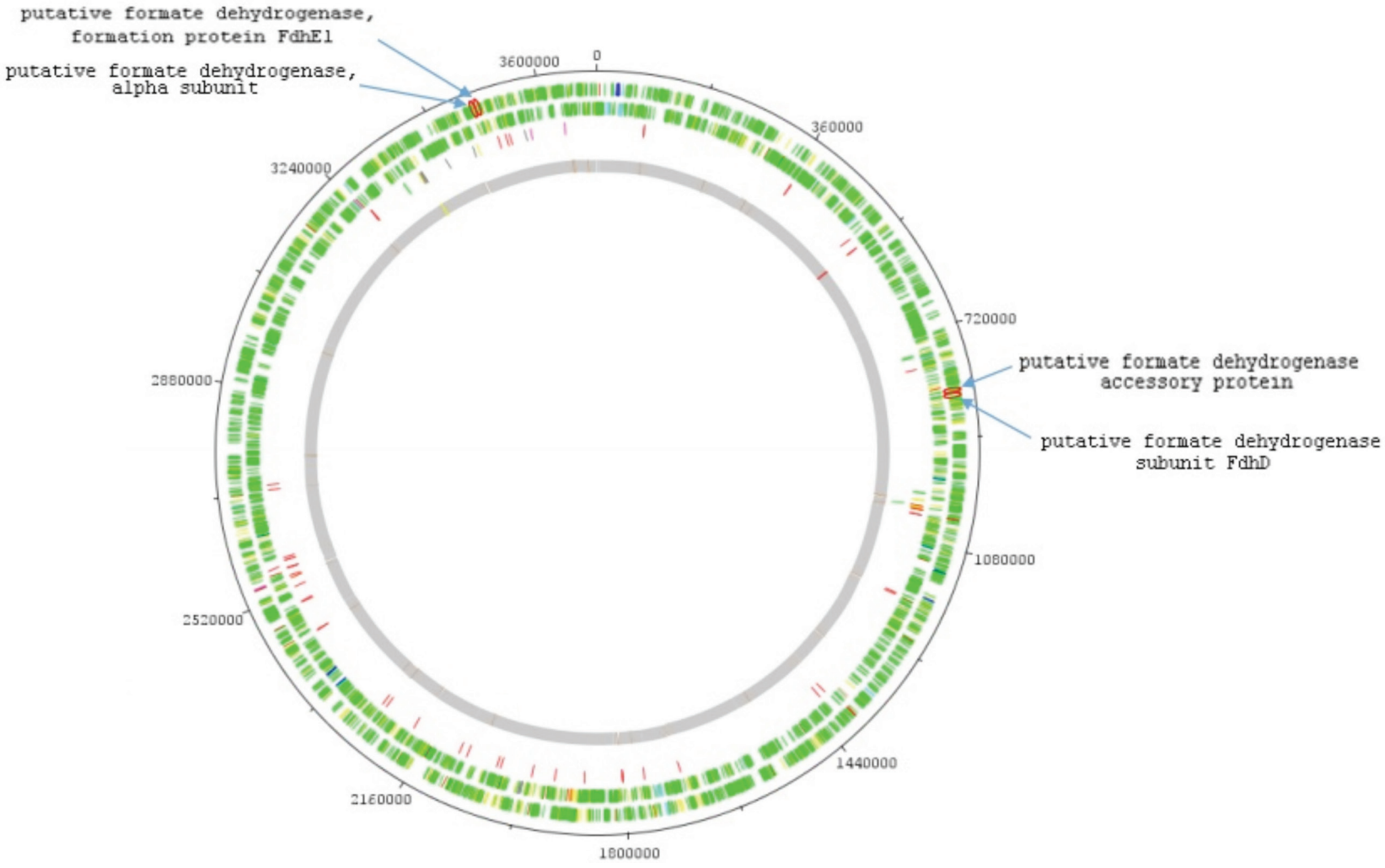


Figure 68 - *Localization of the genes coding for tugnsten containing enzymes on the annotated genome, obtained from artemis software.*

Chapter V

Preface

This chapter is dedicated to trends in the genomics world, and the direction that NGS is taking. We are in 2014, and both the BIG DATA output and cost per Kb of genomics information could not be predicted, not even in 2012, only two years ago. It is impossible to predict how the genomics world will be in 2020, only six years ahead.

This last chapter, besides presenting interesting numbers and graphics, collected from very recent economic surveys targeting investors, is also about the importance of translational research.

In fact, the *Desulfovibrio gigas* genome project, after all, is not only a scientific project, aimed purely at knowledge. Because of the partnership between ITQB and STAB VIDA, the company born in July 2001, this genome is also a successful story of industry-academia partnership to generate knowledge and economic value for society. A few take home messages are suggested and a possible model for boosting translational research and its benefit for Portugal's economic growth is presented. Such model, called 5.50.500, is suggested also as a way for Portuguese genetics professionals contribute to the re-balancing of ERA (European Research Area).

Acknowledgments: the work presented in this chapter represents mostly the author's point of view. Daniela Leão and Sofia Goes, shareholders of STAB VIDA, are to be acknowledged for their work of daily debate and in practice of the concepts of translational research hereby discussed.

"*Things don't have to change the world to be important*"

Steve Jobs

V.1 -The evolution of BIG DATA output overtime, and what is next on the corner.

The field of Life Sciences is, more and more, relying on BIG DATA produced by the NGS technologies. Many different applications exist today, being small genome sequencing a fraction of these. At STAB VIDA we receive many requests of quotation and projects that researchers outsource to specialized companies like ours, being the number of NGS projects growing exponentially, as shown in Fig. 69. For some of these projects, five million reads with average size of 300 bp is enough, for some others, even 20 million reads are not enough. One of the major bottlenecks is still the bioinformatics analysis, where for such quantity of data, better algorithms are needed, and interpretation of the results is not yet user friendly.

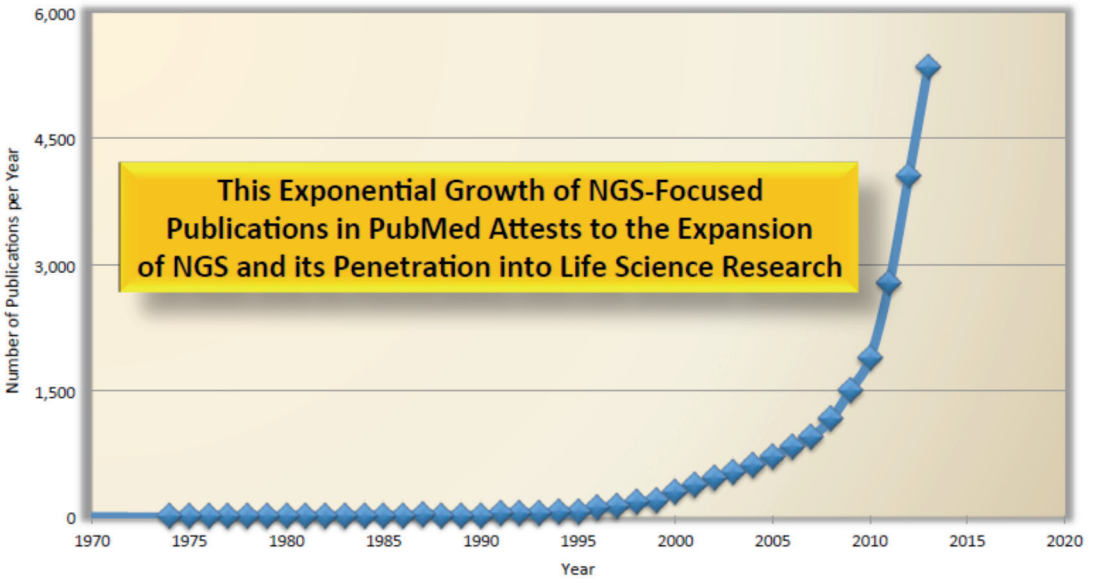


Figure 69 – *Focused Publications using BIG DATA in Life Sciences.*

Table 36 contains a summary of the explosive evolution of the output in terms of base pairs per day, from the 70's till today.

Table 36 – Milestones in DNA sequencing technology.

	Output (base pairs per day)
1970's Manual sequencing using radioactive isotopes for tagging DNA	1,500
1985 First Automated Sequencer ABI model 370A DNA tagging DNA	6,000
1990 ABI PRISM Model 373 DNA Sequencer	9,600
1995 First Capillary Electrophoresis Sequencer (ABI PRISM 310)	15,000
1995 ABI Prism 377 DNA Sequencer	105,000
1997 MegaBACE 1000	500,000
1998 PE Biosystems Prism 3700	1,000,000
2001 MegaBACE 4000	2,800,000
2005 First 'Next Gen Sequencer" GS20 454 Life Sciences	20,000,000
2007 Illumina Genome Analyzer	150,000,000
2009 Genome Analyzer IIx and SOLiD 3	5,000,000,000
2010 HiSeq 2000	25,000,000,000
2012 HiSeq 2500	90,000,000,000
2014 HiSeq X Ten	600,000,000,000

The following sections summarize interesting conclusions taken by recent surveys and market analysis targeting NGS consumers. These data were gathered in the following surveys:

i) Razvi E (2014) Next Generation Sequencing (NGS): Market Trends. GENReports: Market & Tech Analysis, Selected Biosciences Inc.,

ii) Groberg J (2014) DNA Sequencing - Genomics 2.0: It's just the beginning. Macquarie (USA) Equities Research, Macquarie Capital (USA) Inc., USA.

iii) UK.Marketsandmarkets.com (2014) Next Generation Sequencing (NGS) Market by Platforms (Illumina HiSeq, MiSeq, HiSeqX Ten, NextSeq 500,Thermo Fisher Ion Proton/PGM), Bioinformatics (Exome Sequencing, RNA-Seq, ChIP-Seq), Technology (SBS, SMRT) & by Application (Diagnostics, Personalized Medicine) – Global Forecast to 2020. MarketsandMarkets (report code: BT 2697), USA.

V.2 -What are the preferences of NGS users?

Illumina platforms are the most widespread of all NGS technologies and the trend for bioinformatics available options are the open-source instead of commercial solutions. As can also be seen from the Figure 70 till Figure 73, the satisfaction of users with NGS output and cost is good (Razvi, 2014).

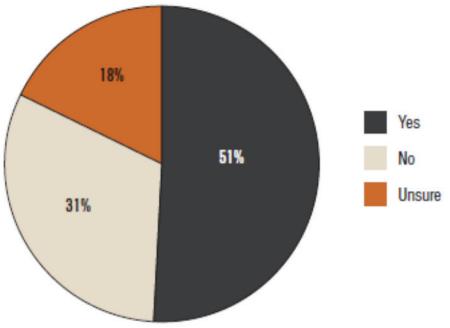


Figure 70 – *Are you satisfied with the lowered cost, higher output, and integrated offerings coming from NGS platform products today?*

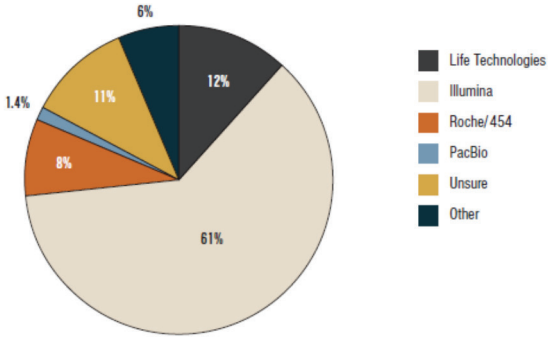


Figure 71 – *Do you own, or support one of the following NGS platforms?*

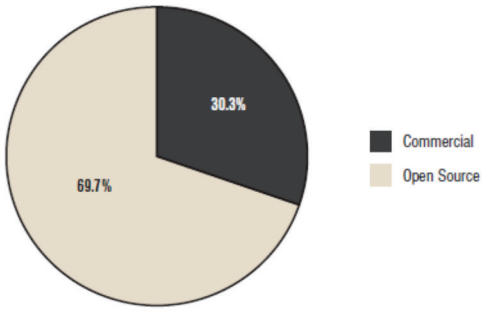


Figure 72 – **Does your organization prefer commercial or open-source NGS software solutions?**

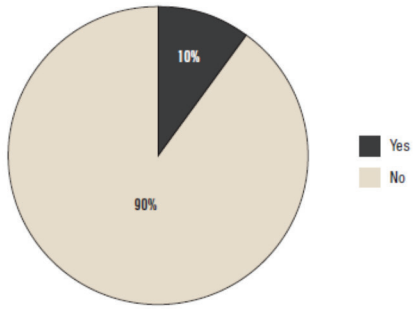


Figure 73 – *Are you currently planning to outsource NGS?*

V.3 – What is the total market size for NGS? The evolution of the costs of DNA sequencing over time: from 15€ for 800 bp in 2001 to 1,5 € for 1 Mb in 2014.

NGS is on a growth trajectory and the NGS-based revenues are expected to reach US $5 Billion by 2015 (Groberg, 2014). The worldwide installed base of NGS instruments for Research Use Only (RUO) in 2014 is forecast at 5,500. For clinical use the number of NGS instruments forecast for 2014 is 1000.

In 2001, when STAB VIDA was initiated, the cost of sequencing the human genome was 100 Million dollars (approx. 75M€). Today, STAB VIDA is offering small bacterial genomes for around 800€ and human exomes for around 900€. It will be obviously very difficult to predict the prices for 2020.

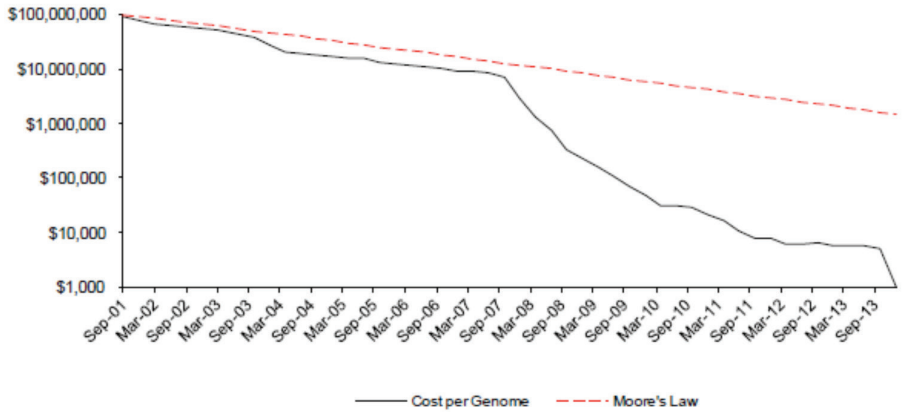


Figure 74 - *The fall in cost per genetic data point has seemingly outpaced Moore's Law*

The figures taken from different surveys (Razvi, 2014; Groberg, 2014; Markets & Markets, 2014) allow us to have an overview on interesting numbers, namely the existence of 50,000 molecular biology labs, out of which 8 are major sequencing centres: the size of NGS market is distributed among different target areas, where molecular bacteriology accounts today for an important fraction, with tendency to grow further.

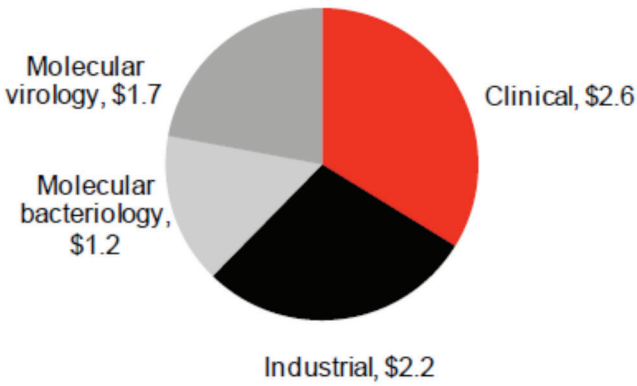


Figure 75 - *Microbiology market size - Microbiology addressable market (in billions of dollars).*

Table 37 - Number of labs globally

	Approximate number of labs
Total laboratories	200,000
Molecular Biology labs	50,000
Clinical with molecular capabilities (US)	2,300
Forensics with molecular capabilities (global)	1,000
Estimated number of labs with at least one sequencer	5,000

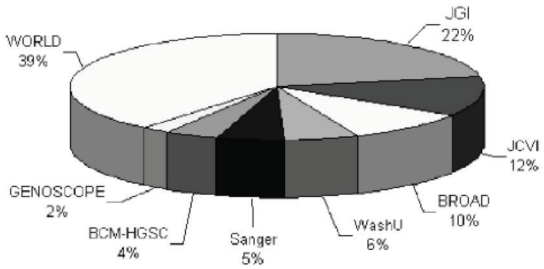


Figure 76 – *Major Sequencing Centres, September 2009 – The most Major Sequencing Centres are located in USA.*

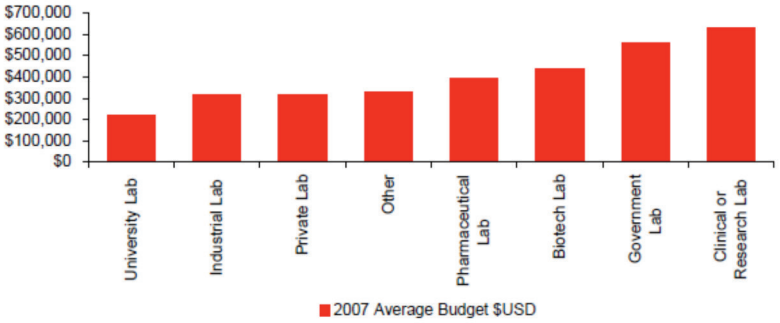


Figure 77 - *Average lab budget by industry pre-recession*

Figs 78 and 79 summarize the panorama of NGS options, in 2011 and 2014, where it is visible an extraordinary evolution

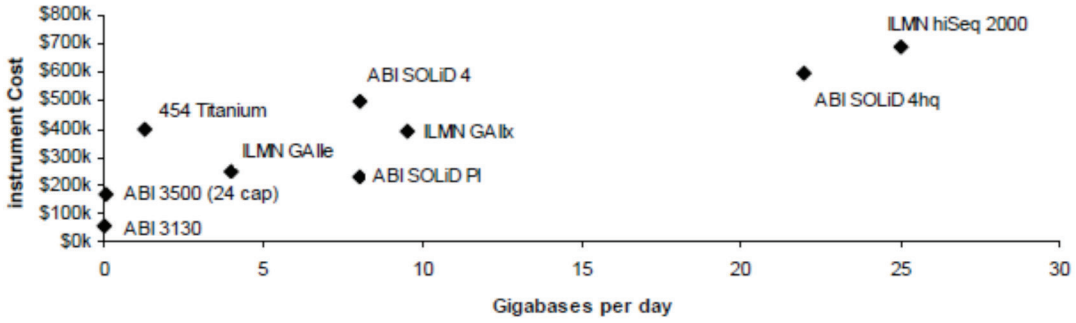


Figure 78 - *Sequencer instrument cost versus output per day, 2011.*

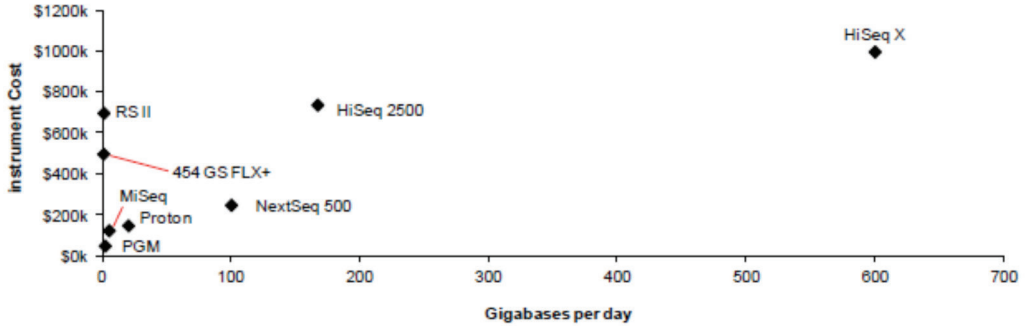


Figure 79 - *Sequencer instrument cost versus output, 2014.*

Table 38 gives some clues on the market size in terms of millions of dollars, for Genomics and NGS.

Table 38 - Total addressable Genomics 2.0 market opportunity (Ex-USA refers to markets outside USA).

(US$m)	USA	Ex-USA	Total
Total Oncology	**$5,600**	**$5,600**	**$667,899**
Clinical	$3,700	$4,600	$8,300
Hereditary	$1,400	$400	$1,800
Molecular monitoring	$500	$600	$1,100
Total Reproductive	**$1,884**	**$4,670**	**$6,554**
NIPT	$950	$1,900	$2,850
Carrier	$500	$1,200	$1,700
Neonatal	$200	$1,000	$1, 200
IVF	$234	$570	$804
Total other markets	**$2,430**	**$2,480**	**$4,910**
Microbiology	$1,500	$1,500	$3,000
Livestock and agriculture	$400	$420	$820
Transplant/HLA	$160	$330	$490
Forensics	$270	$150	$420
Consumer genomics	$100	$80	$180
Total addressable market opportunity	$10,000	$13,000	$23,000

V.4 – Who is who in genomics? And what is genomics important for? What is the importance of small genomes for professionals, in particular, and for science as a whole?

Next table shows the major players investing in NGS platforms, as seen worldwide. In the Iberian Peninsula, STAB VIDA is already a strong player, by operating the platforms Miseq, PGM and Hiseq 1500.

Table 39 - Most promising NGS companies

Top 20 Genome Centres	Number of 2nd gen instruments
Broad Institute	109
BGI (formerly Beijing Genomics Institute)	128
Ignite Institute	100
The Genome Center at Washington University	58
Wellcome Trust Sanger Institute	42
DOE Joint Genome Institute	20
Baylor College of Medicine	18
Michael Smith Genome Sciences Centre	16
Ontario Institute for Cancer Research	15
Beijing Institute of Genomics	15
Centro Nacional de Analisis Genómico (CNAG)	10
Genome Institute of Singapore	9
Cold Spring Harbor Laboratory	8
Centre for Genomic Research	7
Beckman Coulter Genomics (formerly Agencourt)	6
UCL Genomics	6
JCVI	6
Cambridge Research Institute	5
GATC	5
Duke IGSP Sequencing Core Facility	5
STAB VIDA	3

A summary of the centres being the most active players in the NGS fields, as well as their major interests is shown in fig 80-83. The most promising companies and players of the future are also presented in Table 39.

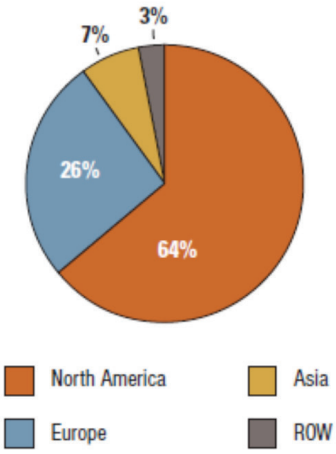


Figure 80 - *Study Demographics*

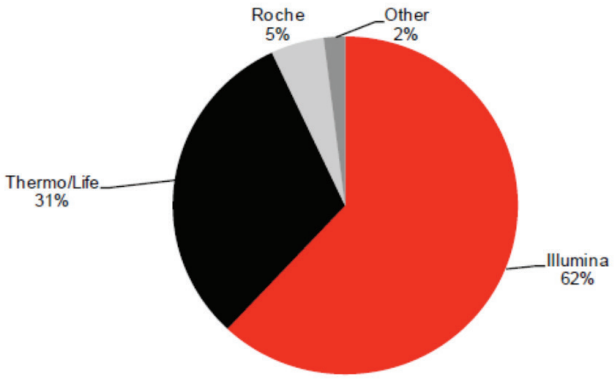


Figure 81 – *Sequencing platforms installed in the market, as visible Illumina accounts for 62% of the installed sequencing platforms in the market.*

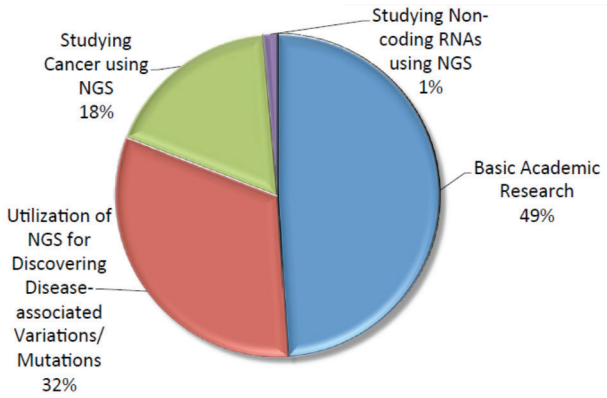


Figure 82 – *Breakout of Researchers vis-à-vis their Utilization of NGS.*

Figs. 83 and 85 show that here is a the focus on oncology but other disease classes are gaining traction too. An important note is percentage breakout between activities focused on analysing somatic mutations versus RNA-Seq.

Taken together, we believe that the data presented herein provide a picture of the NGS field, and furthermore we are seeking to provide starting material for NGS-based panel tests which can then be validated across patient cohorts and developed subsequently into LDTs/IVDs.

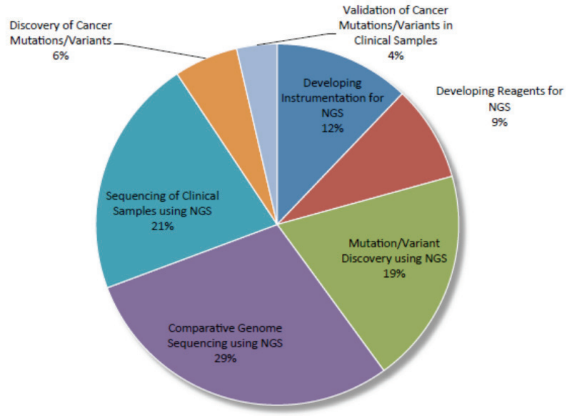


Figure 83 - *Breakout of Researchers vis-à-vis Their Utilization of NGS. II*

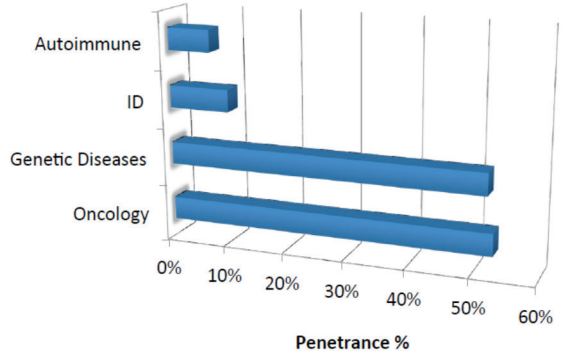


Figure 84 – *Segmentation of the NGS clinical Space by Disease Class Addressed currently*

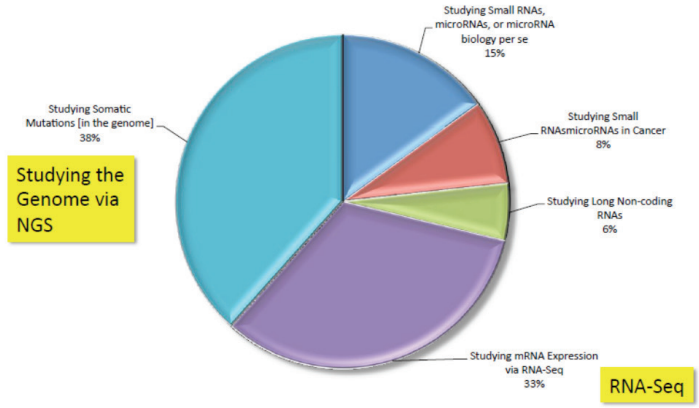


Figure 85 - *"Analyte Classes" Studied via NGS Today: Provides a Picture of Research Efforts by Type of Nucleic Acid*

V.5 – The importance of small genomes sequencing

NGS is surely an important tool for the study of microorganisms and their effects on other living organisms with an emphasis on infectious-disease causing bodies which is of particular interest to public health scientists. There are three main buckets into which microbiology can fall: basic research, clinical, and industrial.

Microbiology basic research

Public health scientists are constantly looking for new methods to understand how diseases are spread in populations to better control future outbreaks. Historically, Sanger sequencing, Multilocus Sequence Typing (MLST) and Multilocus VNTR Analysis (MLVA) have been the preferred methods for public health departments to research outbreaks. Time, money, and labour resources needed for these methods have become the limiting constraints. (Goering *et al,* 2013)

Over the past few years, however, NGS has taken over for prior processes as scientists can now sequence thousands of bacterial isolates affordably and cost efficiently while looking at an entire genome rather than just bits and pieces.

With the high throughput capabilities of NGS, pathogen samples from multiple patients can be quickly sequenced to identify mutational markers and ultimately understand the social relationships and pathways through which diseases spread. Basic research is essential to pave the way for the future method for identifying these highly infectious pathogens (that can quickly spread without proper identification).

Clinical microbiology

In hospital settings, being able to determine the identity and anti-biotic susceptibility of infectious organisms is critical. "Classic" microbiology involves seeding Petri dishes with specialized media and seeing which organisms grow. Then, based on an efficient isolation of a bacteria colony in the dish, the organism can be tested for antibiotic susceptibility. While highly effective, this process can take days to weeks to deliver results, and patients are often started on an antibiotic regimen as a precaution. The reasons classic microbiology persists today are its cost (10x less than molecular testing) and its definitive guidance for antibiotic prescription. However, this will change over time.

As a first step in this transition, many labs are adopting new mass spectrometry based systems in order to more quickly and cost effectively ID microorganisms. A leader in this field is Bruker (BRKR) with its MALDI Biotyper. With the MALDI Biotyper the lab must still grow the organism, but instead of the cumbersome steps to identify the organism they use spectra generated from the mass spec. This shaves ~24 hours off the typical time-frame.

In the future, most microbiology labs think that all these organisms will simply be sequenced.

As such from sequence information, the labs will be able to know not only the bug's identity but also its susceptibility. However to get to that point, the cost of library preparation must get closer to the costs of classic microbiology, or ~€10. Today NGS library preparation is on average ~€30-35 (depending on the volume of the organisms a lab might have), but this is a cost that continuously will fall quickly. The demand for infectious disease identification should grow as far as the prices decline further and the cost effectiveness of NGS applications become fully validated.

Industrial microbiology

Industrial applications for microbiology include testing food, water, animals, or manufacturing processes (e.g. biologics) for infectious organisms. Industrial applications tend to be much more price sensitive, but remain an important part of the microbiology market.

Sizing up the microbiology market

Although likely a much longer term opportunity, the estimated market size could be near €2.6 billion in clinical microbiology, €2.2 billion in industrial, €1.2 billion in molecular bacteriology, and €1.7 billion in molecular virology for a total addressable market size of ~€8 billion (Groberg 2014; Markets & Markets, 2014).

NGS, over time, could replace many of the existing technologies used in microbiology. If one assumes that 1/3 of the ~€8.0 billion microbiology market converts to NGS over time, this could represent a €3 billion market opportunity (Razvi, 2014; Groberg, 2014).

V.6 – The project Gigasnoma for solving the *D. gigas* genome sequence is an industry-academia partnership and is a case study of the difficulties of translational research

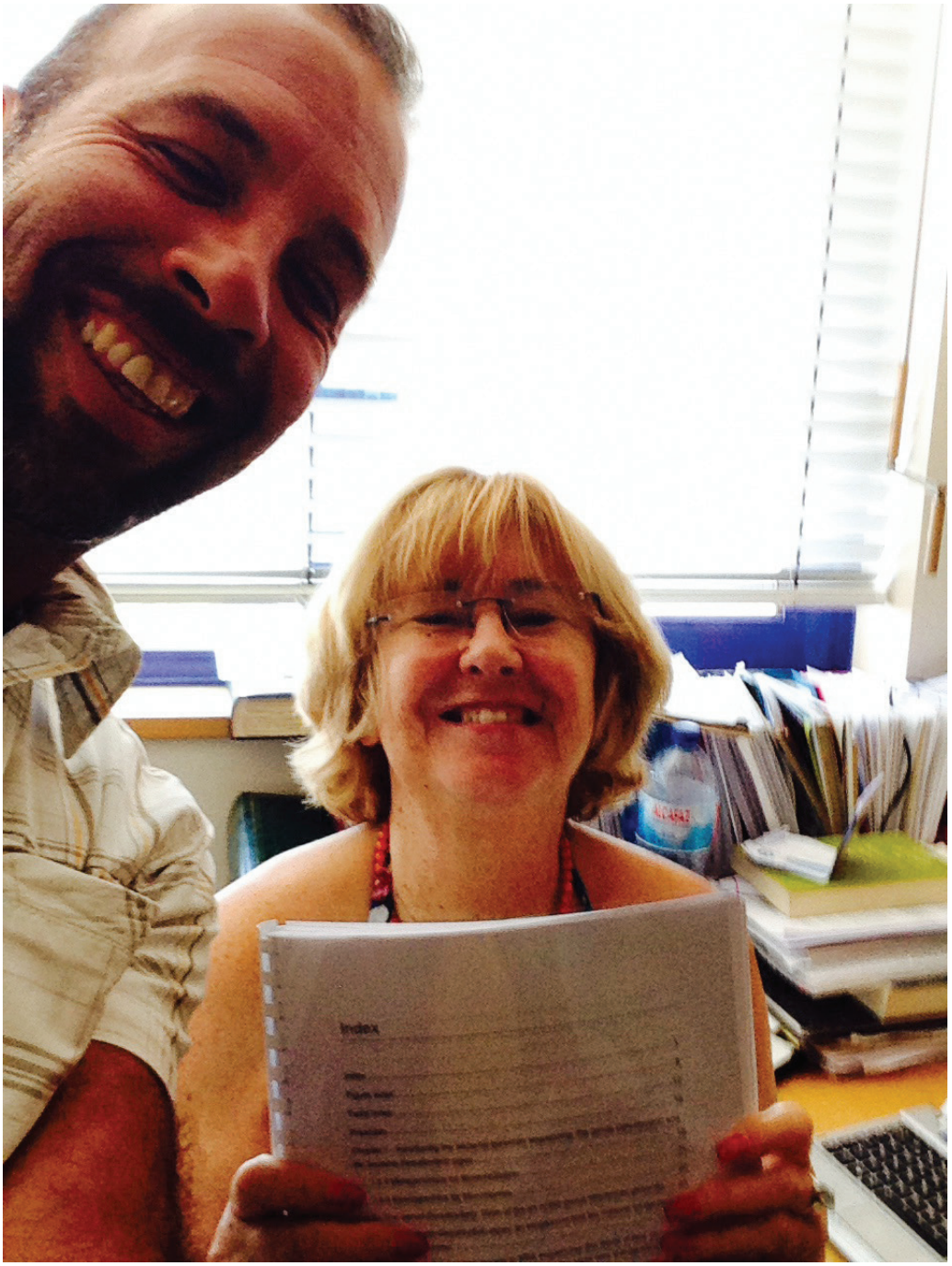


Figure 86 - First draft of this thesis, delivered to Prof. Rodrigues-Pousada, 15th July 2014. Both the author and Prof. Rodrigues-Pousada are the faces of this industry - academic collaboration for translational research.

The work of solving the *D. gigas* genome was done in parallel with the very absorbent tasks of launching and running a biotech start up that, coincidentally (or not) performs DNA sequencing as one of its core activities. Looking back, now I can say that the benefits were in both directions: from *D. gigas* project to STAB VIDA service implementation, and from the sequencing routines implemented at STAB VIDA, for delivering more and more data to the *D. gigas* genome project.

Gigasnoma can be considered a case study for translational research as a result from the "learning by doing" process of industry-academia partnership between ITQB and STAB VIDA. This partnership was one of the pioneering initiatives in Portugal for translational research and mutual benefit for the two sides of the same coin: industry and academia.

V.7 Why translational research? Knowledge creation: it's nice, but..... it is also an investment that needs return.

Knowledge is created out of investment (money and time), mostly making use of public money, meaning tax payers money. It is only fair to expect that knowledge should pay this investment, back to society.

In life sciences, expenditure is very high, and there are Institutes where it costs an average of 250k€ per PI per year, plus his/her salary. Such amount of investment generate knowledge from data, for being published.

If such knowledge will generate new products or new services, or important support for correct decision making (for instance in the clinical practice), then the investment will revert again to the tax payers in the form of:

i) new collected taxes from the new products, services and/or

ii) improved healthcare, improved food or improved quality of life in general, etc.

But the process of knowledge translation is almost a science *per se*, it is very hard to achieve success, very intensive and not at all simple. Nevertheless, it has to be encouraged since generating knowledge without translational application will not outreach and will not add benefit for society as a whole.

One successful example is the genome project of *D. gigas*. Solving and annotating the complete genome is the evidence that knowledge has been generated. The Portuguese tax payers contributed to it with a 100 k€ investment, through a grant of AdI to ITQB (80%) and STAB VIDA (20%). On plus, STAB VIDA invested from own private funds around 150 k€. But after the conclusion of this project, and also because of it, STAB VIDA became international service provider of small genomes sequencing and BIG DATA/bioinformatics selling, returning more than those 100 k€ in taxes to the tax collector, each year, and with good perspectives of growing the overall revenues of the company.

Or, saying in another way: with a society investment of total 100 k€, the same society now gets more than those 100 k€ back every year, and 3 new jobs have been created – all that plus, of course, the knowledge of the full annotated genome of the bacteria *Desulfovibrio gigas,* openly accessible by anyone, anywhere in the world.

There are however many cases of investments to obtain new knowledge that are not translated at all. We all know that, and there is nothing wrong about it, as long as the scientists do understand that, whatever knowledge they produce is sponsored by tax payers, and they are expecting a benefi-

cial return to the whole community. Researchers should be encouraged to cooperate in the process of translation, probably not being the drivers of it – it takes the right professionals for that – but being participants and direct beneficiaries of that.

Knowledge creators are the professionals that society chooses for investing on, hoping to get back a good ROI (return on investment). Innovators are the professionals needed to make it a reality. Knowledge is surely not more than 50% of it. Translation is the other 50%. Society needs both working together for translating the research, and both of them need each other.

V.8 The reality of translational research in Europe: the continent that lives the peace and unbalance.

Europe is a Continent in peace, but unbalanced. The 28 EU members are enjoying decades of peace and progress. Such progress, however, is unbalanced, since the Southern periphery of the EU 28 is poorer and struggling with unemployment and bankruptcies, while the core Northern countries have superavits on their trade balance, low unemployment rates and solid economic growth.

Keeping up peace was probably the most important initial driver for the European Union idea and project. Now, the times are different, and there are other big challenges and threats to the Europeans and to the EU Member States. Major risks are energy supply and unbalance. If one would be able to look from the sky over EU-28, a huge flow of money in the form of interests of the State's debt, and skilled professionals (immigrants) is seen from some of the peripheral countries (poorer and less industrialized, less specialized economies) to the northern core EU (most innovative economies, richer, high earning in interests from near-past loans to the poorer countries).

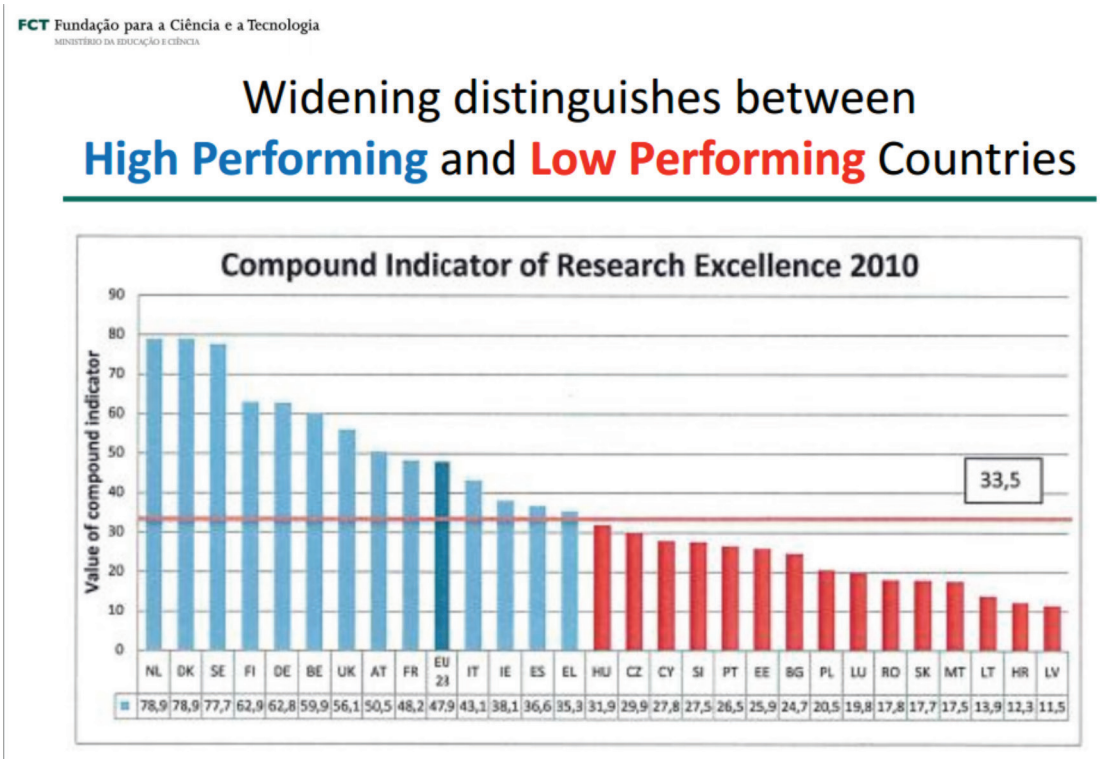


Figure 87 - Portugal and other peripheral countries were performers in the EU FP7 framework program. This means that the country's contribution to EU budget was higher than its re-attracted budget.
Low performing countries are liquid contributors to high-performing countries, like DE and UK.

These countries have, *a priori*, two advantages over other countries like Portugal, Bulgaria, Slovenia, etc.: the language and the good reputation that involuntarily influences the evaluators when ranking the applicants and their proposal. These framework programs, somehow contribute to the depletion of resources (financial, young skilled people) in the periphery towards the core. This has a huge impact in the scientific landscape of the European Research Area (ERA), where peripheral countries are liquid contributors to the core EU countries, which are liquid receivers.

In the end, ironically, there is an indirect and direct support of the progress in the form of better research and innovation of the richer countries, from the poorer. Consequently, years after, these innovations, meanwhile fabricated in the richer countries, will be bought and imported from the peripheral countries at very high prices, the same innovation they helped financing, resulting in a tremendous dependence in terms of their freedom to operate, running costs, etc., turning it into a vicious cycle.

European Research Area needs urgent balance. ERA is fundamental for the harmonious growth of all EU countries, and vital for the less developed regions in the periphery. Our politicians, and other decision makers, need to play their role and fight for the implementation of a harmonious ERA. An unbalanced Europe will not last, will disintegrate and will affect its citizens. Above all it is not morally acceptable.

To generate balance, i.e., to re-equilibrate the ERA, research and innovation needs to be developed and retained in the periphery. This can only be achieved if a cluster of innovative professionals and innovative organizations start developing and growing near the fertile centres of research, like the Universities and Institutes of Research.

Universities from less developed countries alone cannot rebalance ERA, because their freedom to hire, freedom to operate and freedom to decide is very limited due to their dependency on the State, due to the immense restrains they face. On the contrary, none of these restrains apply to innovative SMEs, where, for instance, doubling its working staff in one year is not as rare as one could expect.

The European Research Area (ERA) is, like EU28, unbalanced. The future of our field professionals (scientists, innovators, technology entrepreneurs) depends on its urgent rebalancing and this can only be achieved if academics and private professionals work closely together, like ITQB and STAB VIDA did for the D. gigas genome. For that the number and impact of contracts between both parts have to rise significantly over the next years, and not longer than 2020.

V.9 – The lack of tradition in academic-industry partnering in Portugal – is our country lagging behind?

In Portugal, as in other countries, private companies are big contributors of taxes to society (IRS, IRC, VAT, social security, and other taxes). But society is suspicious of all managers of private companies without knowing that, if companies fail in delivering their taxes, their manager's assets can, by law, be sold to cover the company's responsibilities, and these managers can be held criminally responsible.

Society takes a lot of benefits from having healthy private companies. Is the reverse valid also?

For instance, let's think about a research institute that is fully equipped with state of art technologically advanced machinery, bought with tax payers money or with debt: what would happen if a start-up company approaches and asks to use such equipment for their commercial purposes? My experience is that all kind of barriers will be built: build all kind of barriers bureaucratic, financial, administrative until the SME finally gives up.

Why such assets, cannot be acceded for a purpose of economic exploitation, in order to potentiate and generate return on investment? Is it bad or good that an investment of, for instance, 1M€ done by the community, primarily for research, a non-commercial purpose, is exploited also by private agents of commercial purpose, taking into account that, the 1M€ came from these category of agents? And what would be the answer if society becomes aware that the parallel use of such asset, if successfully done, can generate an income of another 1M € for the tax authorities (IRS, Social Security, VAT, IRC, etc)?

Now, lets think on the investment that the Portuguese society has done in recent years in libraries, equipments, buildings, P1, P2, P3 laboratories, software, hardware, cold chambers, fermentors, bioreactors, scanners, sequencers, mass spectrometers, RMN, etc, etc. Would it be reasonable to say that the whole BIO academic/scientific system in Portugal operates on top of an installed and payed investment of 1.000 Million of Euros, coming from the tax payers (and loans, to be payed with interest)? How much sales do all these assets generate? How much tax society gets back from their investment? 1 M€ per year? 10 M€ per year?

Portugal and all the Portuguese citizens have the task of looking to all that exists as exploitable assets, and pave all possible ways for taking full advantage of them. It is a professional mission but above all, it is an urgency for a country in huge deep crisis.

V.10 - Suggesting a model for fostering growth involving effective translational research

The model 5-50-500-5000 or 50 SMEs – 5% royalties

The following model is a suggestion for encouraging translational research to achieve a fast growth and seriously contribute to the Country's GDP growth as well as the ERA's re-balancing. In this model, the most knowledgeable centres must produce:

i) knowledge,

ii) skilled human resources and

iii) economic activity, through hosting SMEs for new commercial products and services.

So far, the knowledgeable centres only produce the first 2, but not the last: economic activity.

In this proposition, a University campus with 5000 graduate students and about 500 PhD students is encouraged to host 50 technological SMEs, who will run upon 500 private jobs and give back to the campus 5% of their revenues. This population of 50 SMEs will be renewed,

at the same time some of them are spanned out, and exit to the neighbouring private parks of science and technology.

How would it work in practice? Let's imagine a scientific campus with 5000 undergraduate students and its staff of Professors, researchers, technical and administrative supporting personnel, with a yearly spending budget of 50 M€. What could change in the years to come, and be fully implemented by 2020:

A - This campus should find enough space (offices and labs) for hosting 50 SMEs, covering different areas (e.g. Life Sciences, materials, Health related companies, medical devices, internet development, bioinformatics, etc.). The attraction of SMEs can be done via advertising (locally, regionally, nationally and abroad) or by stimulating an internal entrepreneurship program for spinning out new companies from the campus existing academic laboratories. As attraction factors, the Directors of these campus could think of minimum rental costs for m2 (between 5 and 10 €/m2), *policy of minimum pay-per-use* of existing infrastructure and installed equipment, defining 2nd priority of commercial exploitation of assets as a universal rule, and sharing of structure like internet, water and gases.

B - The managers of these companies and the PIs of the campus should, to the best possible allocation, have their office spaces in common (mixed) office rooms. The same should happen between the sitting desks of PhD students and the staff of those SMEs.

C - Each SME should as a universal rule give 5% of their yearly sales to the Academic Campus, independently of their profit. And when their turnover surpass the 5 M€, again independent of profit, these spin off and startups need to exit the campus, and be transferred to a neighbouring Science Park.

On average, these 50 SMEs would have 10 staff members each, and contribute to the State in taxes with 1 M€ yearly, each. This means 500 private jobs (some of them direct employment of human resources trained in the *campus*) and a full self-sustainability of financials, since the State's contribution budget would be 50 M€ to the knowledge campus but it would collect the same 50M€ in taxes from the 50 SMEs the same amount of 50 M€. On plus, an employment structure (trainees, first jobs, etc.) would be living on campus door to door with the training laboratories and classrooms.

Once the SMEs exit to the Science Parks, they are more robust and their chance of survival, growth and exports generator is already high. If they decide to stay 5% contributors to the academic campus, this value is deducted from their IRC. A successful campus would spin out 3 to 5 SMEs each year, and keep from these 5% royalties of sitting SMEs plus 5% of alumni SMEs, around 5M€ every year, that the campus could use for funding internal research programs.

The university campus and the research centres, in Portugal, are a kind of "sleeping beauties" and they need to engage proactively in a national program of translational research, rendering direct benefits for them, but also huge benefits for society.

V.11 – 20 years.... looking back and looking ahead

From the first paper of this thesis (Ulrich et al, 1994), till the 2nd and most important one (Morais-Silva *et al*, 2014), 20 years have passed.

This was a long time.

Looking back, one must say that the genetics field had a revolution, moving from gene studies that could take 1 year or more to whole genome studies that only need one week maximum. The amount of data generated today is astonishing.

In the quest of justification for the need of so much time for the presentation of this work, many different reasons came upon as real. Like for instance the technological evolution, the bioinformatics bottleneck, the demanding job of managing a start-up. But, in the end, the most obvious explanation is one: time is too fast! Really fast! But, above all, it was a great team.

What is the future? Most probably whole microbial genomes and whole human exomes will become daily routines in any genetics lab. But if I am allowed to do a futurology exercise, I would dare the following previsions, but please only to be read by 2050:

i) The organization of society and cities will be completely different: in the recent past, churches were substituted by shopping malls as the main society gathering places. In the future, shopping malls will be built underneath the cities and they will supply food and daily needs from below, by elevators, directly to the homes of citizens, after ordering on the internet (just like water and electricity supply).

ii) People will live longer, and each city hospital will have an extension at every one's home, with point of need kits serving as pre-diagnostics routine procedure. There will only be 5 countries left after territorial re-organization.

iii) Someone will prove that the major role of DNA in the cell and in the living organisms is not coding for proteins nor hereditary related functions. These are probably the secondary effects, of the major role of DNA in life yet to be proved...

Bibliography

- Agostinho, M., Oliveira, S., Broco, M., Liu, M. Y., LeGall, J., & Rodrigues-Pousada, C. (2000). Molecular cloning of the gene encoding flavoredoxin, a flavoprotein from *Desulfovibrio* gigas. *Biochemical and Biophysical Research Communications*, *272*(3), 653–6.
- Barton, L. L., and Fauque, G. D. (2009) Biochemistry, physiology and biotechnology of sulfate-reducing bacteria. *Adv. Appl. Microbiol.* 1st ed., pp 41–98. Elsevier Inc.
- Bell, T. H., Joly, S., Pitre, F. E., and Yergeau, E. (2014) Increasing phytoremediation efficiency and reliability using novel omics approaches. *Trends Biotechnol. 32*, 271–280.
- Borges, V., Ferreira, R., Nunes, A., Sousa-Uva, M., Abreu, M., Borrego, M. J., & Gomes, J. P. (2013). Effect of long-term laboratory propagation on Chlamydia trachomatis genome dynamics. Infection, Genetics and Evolution : *Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 17, 23–32.
- Broco, M., Rousset, M., Oliveira, S., & Rodrigues-Pousada, C. (2005). Deletion of flavoredoxin gene in *Desulfovibrio* gigas reveals its participation in thiosulfate reduction. *FEBS Letters*, *579*(21), 4803–7.
- Buchholz, U., Bernard, H., Werber, D., Böhmer, M. M., Remschmidt, C., Wilking, H., Deleré, Y., an der Heiden, M., Adlhoch, C., Dreesman, J., Ehlers, J., Ethelberg, S., Faber, M., Frank, C., Fricke, G., Greiner, M., Höhle, M., Ivarsson, S., Jark, U., Kirchner, M., Koch, J., Krause, G., Luber, P., Rosner, B., Stark, K., and Kühne, M. (2011) German outbreak of *Escherichia coli* O104:H4 associated with sprouts. *N. Engl. J. Med. 365*, 1763–70.
- Chan, E. Y. (2005) Advances in sequencing technology. *Mutat. Res. 573*, 13–40.
- Dark, M. J. (2013) Whole-genome sequencing in bacteriology: state of the art. *Infect. Drug Resist. 6*, 115–123.
- Doyle, J. M., Katzner, T. E., Bloom, P. H., Ji, Y., Wijayawardena, B. K., and Dewoody, J. A. (2014) The Genome Sequence of a Widespread Apex Predator, the Golden Eagle (Aquila chrysaetos). *PLoS One 9*, e95599.
- Fareleira, P. (2003) Response of a strict anaerobe to oxygen: survival strategies in *Desulfovibrio* gigas. *Microbiology 149*, 1513–1522.
- Farrer, R. A., Kemen, E., Jones, J. D. G., and Studholme, D. J. (2009) *De novo* assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads. *FEMS Microbiol. Lett. 291*, 103–11.
- Félix, R., Rodrigues, R., Machado, P., Oliveira, S., and Rodrigues-Pousada, C. (2006) A chemotaxis operon in the bacterium *Desulfovibrio* gigas is induced under several growth conditions. *DNA Seq. 17*, 56–64.
- Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., and al., e. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science 269*, 496–512.
- Goering, R., Köck, R., & Grundmann, H. (2013). From theory to practice: molecular strain typing for the clinical and public health setting. *Euro Surveill 2013*: 18(4):20383
- GOLD database V5 (31.05.2014): https://gold.jgi-psf.org
- Gomes, C. M. (1997). Studies on the Redox Centers of the Terminal Oxidase from *Desulfovibrio* gigas and Evidence for Its Interaction with Rubredoxin. *Journal of Biological Chemistry*, *272*(36), 22502–22508.

- Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V, Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. (2010) A draft sequence of the Neandertal genome. *Science 328*, 710–22.

- Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., Rothberg, J. M., Paunovic, M., and Pääbo, S. (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature 444*, 330–6.

- Green, R. E., Malaspinas, A.-S., Krause, J., Briggs, A. W., Johnson, P. L. F., Uhler, C., Meyer, M., Good, J. M., Maricic, T., Stenzel, U., Prüfer, K., Siebauer, M., Burbano, H. A., Ronan, M., Rothberg, J. M., Egholm, M., Rudan, P., Brajković, D., Kućan, Z., Gusić, I., Wikström, M., Laakkonen, L., Kelso, J., Slatkin, M., and Pääbo, S. (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell 134*, 416–26.

- Groberg J (2014) DNA Sequencing - Genomics 2.0: It's just the beginning. Macquarie (USA) Equities Research, Macquarie Capital (USA) Inc., USA.
- Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W. J., Wang, X., Xie, B., Ni, P., Ren, Y., Zhu, H., Li, J., Lin, K., Jin, W., Fei, Z., Li, G., Staub, J., Kilian, A., van der Vossen, E. A. G., Wu, Y., Guo, J., He, J., Jia, Z., Ren, Y., Tian, G., Lu, Y., Ruan, J., Qian, W., Wang, M., Huang, Q., Li, B., Xuan, Z., Cao, J., Asan, Wu, Z., Zhang, J., Cai, Q., Bai, Y., Zhao, B., Han, Y., Li, Y., Li, X., Wang, S., Shi, Q., Liu, S., Cho, W. K., Kim, J.-Y., Xu, Y., Heller-Uszynska, K., Miao, H., Cheng, Z., Zhang, S., Wu, J., Yang, Y., Kang, H., Li, M., Liang, H., Ren, X., Shi, Z., Wen, M., Jian, M., Yang, H., Zhang, G., Yang, Z., Chen, R., Liu, S., Li, J., Ma, L., Liu, H., Zhou, Y., Zhao, J., Fang, X., Li, G., Fang, L., Li, Y., Liu, D., Zheng, H., Zhang, Y., Qin, N., Li, Z., Yang, G., Yang, S., Bolund, L., Kristiansen, K., Zheng, H., Li, S., Zhang, X., Yang, H., Wang, J., Sun, R., Zhang, B., Jiang, S., Wang, J., Du, Y., and Li, S. (2009) The genome of the cucumber, Cucumis sativus L. *Nat. Genet. 41*, 1275–81.
- Hui, P. (2014) Next generation sequencing: chemistry, technology and applications. *Chem. Diagnostics*.
- Illumina catalog: http://www.illumina.com/literature.html
- Koboldt, D. C., Larson, D. E., Chen, K., Ding, L., and Wilson, R. K. (2012) Massively parallel sequencing approaches for characterization of structural variation. *Methods Mol. Biol. 838*, 369–84.
- Levy, S. (2013) Ancient gut microbiomes shed light on modern disease. *Environ. Health Perspect. 121*, A118.
- Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O. A., Leung, F. C.-C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C.,

Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C. C., Lam, T. T.-Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M. W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.-W., Yiu, S.-M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G. K.-S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J., and Wang, J. (2010) The sequence and *de novo* assembly of the giant panda genome. *Nature 463*, 311–7.

- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012) Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol. 2012*, 251364.
- Machado, P., Félix, R., Rodrigues, R., Oliveira, S., & Rodrigues-Pousada, C. (2006). Characterization and expression analysis of the cytochrome bd oxidase operon from *Desulfovibrio* gigas. *Current Microbiology*, *52*(4), 274–81.
- Maheshi Dassanayak, D. Oh, J. Haas, A. Hernandez, H. Hong, S. Ali, D. Yun, R. Bressan, J. Zhu, J. M. Cheeseman, and H. J. Bohnert. (2011), The Genome of an extremophile Arabidopsis-relative: Thellungiella parvula. Nature Genetics, 43, pp913–918.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V, Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature 437*, 376–80.

- Marketsandmarkets.com (2014) *Next Generation Sequencing (NGS) Market by Platforms (Illumina HiSeq, MiSeq, HiSeqX Ten, NextSeq 500,Thermo Fisher Ion Proton/ PGM), Bioinformatics (Exome Sequencing, RNA-Seq, ChIP-Seq), Technology (SBS, SMRT) & by Application (Diagnostics, Personalized Medicine) – Global Forecast to 2020*. MarketsandMarkets (report code: BT 2697), USA.
- Matias, P. M., Pereira, I. a C., Soares, C. M., and Carrondo, M. A. (2005) Sulphate respiration from hydrogen in *Desulfovibrio* bacteria: a structural biology overview. *Prog. Biophys. Mol. Biol. 89*, 292–329.
- Mellmann, A., Harmsen, D., Cummings, C. A., Zentz, E. B., Leopold, S. R., Rico, A., Prior, K., Szczepanowski, R., Ji, Y., Zhang, W., McLaughlin, S. F., Henkhaus, J. K., Leopold, B., Bielaszewska, M., Prager, R., Brzoska, P. M., Moore, R. L., Guenther, S., Rothberg, J. M., and Karch, H. (2011) Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PloS one*, 6, issue 7, 1-9.
- Merriman, B., R&D Team, I. T., and Rothberg, J. M. (2012) Progress in Ion Torrent semiconductor chip based sequencing. *Electrophoresis 33*, 3397–3417.
- Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol. 12*, R112.
- Morais-Silva FO, Rezende AM, Pimentel C, Santos CI, Clemente C, Varela-Raposo A, Resende DM, da Silva SM, de Oliveira LM, Matos M, Costa DA, Flores O, Ruiz JC,

Rodrigues-Pousada C., Genome sequence of the model sulfate reducer *Desulfovibrio gigas*: a comparative analysis within the *Desulfovibrio* genus., *Microbiology open*. 2014 Aug;3(4):513-30

- Morais-Silva, F. O., Santos, C. I., Rodrigues, R., Pereira, I. A. C., and Rodrigues-Pousada, C. (2013) Roles of HynAB and Ech, the only two hydrogenases found in the model sulfate reducer *Desulfovibrio* gigas. *J. Bacteriol. 195*, 4753–60.
- Moura, J. J. G., Xavier, A. V., Burschi, M., Le Gall, J., Hall, D. O. & Cammack, R. (1976) A Molybdenum-Containing Iron- Sulfur Protein from *Desulfovibrio* gigas, *Biochem. Biophys. Res. Commun*. 72, 782-789.
- Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res. 39*, e90.
- Nature Publishing Group (2006) Sequencers step up to the speed challenge. *Nature 443*, 258–9.
- Nicole L Quinn, Natasha Levenkova, William Chow, Pascal Bouffard, Keith A Boroevich, James R Knight, Thomas P Jarvie, Krzysztof P Lubieniecki, Brian A Desany, Ben F Koop, Timothy T Harkins and William S Davidson, 2008, Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome, *BMC Genomics* 2008, 9:404
- Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., and Barron, A. E. (2011) Landscape of next-generation sequencing technologies. *Anal. Chem. 83*, 4327–41
- Nikolaki, S., and Tsiamis, G. (2013) Microbial diversity in the era of omic technologies. *Biomed Res. Int. 2013*, 958719.
- Orlando, L., Ginolhac, A., Raghavan, M., Vilstrup, J., Rasmussen, M., Magnussen, K., Steinmann, K. E., Kapranov, P., Thompson, J. F., Zazula, G., Froese, D., Moltke, I., Shapiro, B., Hofreiter, M., Al-Rasheid, K. A. S., Gilbert, M. T. P., and Willerslev, E. (2011) True single-molecule DNA sequencing of a pleistocene horse bone. *Genome Res. 21*, 1705–19.
- Pham N, T. A., and Anonye, B. O. (2014) Vying over spilt oil. *Nat. Rev. Microbiol. 12*, 156.
- Postgate, J., & Kent, H. (1984). The genomes of *Desulfovibrio gigas* and *D. vulgaris*. *Journal of General Microbiology*, 130: 1597–1601.
- Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K., and Wang, J. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature 490*, 55–60.
- Razvi E (2014) Next Generation Sequencing (*NGS): Market Trends*. *GENReports: Market & Tech Analysis, Selected Biosciences Inc., UK.*
- Reuter, S. (2013) Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. *JAMA Intern. 173*, 1397–1404.
- Reuter, S., Ellington, M. J., Cartwright, E. J. P., Köser, C. U., Török, M. E., Gouliouris, T., Harris, S. R., Brown, N. M., Holden, M. T. G., Quail, M., Parkhill, J., Smith, G. P., Bentley, S. D., and Peacock, S. J. (2013) Rapid bacterial whole-genome sequencing to enhance

diagnostic and public health microbiology. *JAMA Intern. Med. 173*, 1397–404.
- Rodrigues, R., Valente, F. M. ., Pereira, I. a. ., Oliveira, S., & Rodrigues-Pousada, C. (2003). A novel membrane-bound Ech [NiFe] hydrogenase in *Desulfovibrio* gigas. *Biochemical and Biophysical Research Communications*, *306*(2), 366–375.
- Rodrigues, R., Vicente, J. B., Félix, R., Teixeira, M., & Rodrigues-pousada, C. (2006). *Desulfovibrio* gigas Flavodiiron Protein Affords Protection against Nitrosative Stress, *J. Bacteriol*., 188(8): 2745-51
- Romão, M. J., Archer, M., Moura, I., Moura, J. J. G., LeGall, J., Engh, R., Schneider, M., Hof, P., and Huber, R. (1995) Crystal Structure of the Xanthine Oxidase-Related Aldehyde Oxido-Reductase from *D. gigas*. *Science 270*, 1170–1176.
- Rothberg, J. M., and Leamon, J. H. (2008) The development and impact of 454 sequencing. *Nat. Biotechnol. 26*, 1117–24.
- Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T., and Bustillo, J. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature 475*, 348–52.
- Roy, A. B., and Trudinger, P. A. (1970) The Biochemistry of Inorganic Compounds of Sulphur. *Cambridge University Press*.
- Schoberth, S. (1973) A new strain of *Desulfovibrio* gigas isolated from a sewage plant. *Arch. Mikrobiol. 92*, 365–368.
- Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., Alkan, C., Kidd, J. M., Sun, Y., Drautz, D. I., Bouffard, P., Muzny, D. M., Reid, J. G., Nazareth, L. V, Wang, Q., Burhans, R., Riemer, C., Wittekindt, N. E., Moorjani, P., Tindall, E. A., Danko, C. G., Teo, W. S., Buboltz, A. M., Zhang, Z., Ma, Q., Oosthuysen, A., Steenkamp, A. W., Oostuisen, H., Venter, P., Gajewski, J., Zhang, Y., Pugh, B. F., Makova, K. D., Nekrutenko, A., Mardis, E. R., Patterson, N., Pringle, T. H., Chiaromonte, F., Mullikin, J. C., Eichler, E. E., Hardison, R. C., Gibbs, R. A., Harkins, T. T., and Hayes, V. M. (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature 463*, 943–7.
- Sekizaki T, Takamatsu D, Osaki M, Shimoji Y., 2005, Different foreign genes incidentally integrated into the same locus of the Streptococcus suis genome, *J Bacteriol.* Feb;187(3):872-83.
- Shenderov, B. A., and Midtvedt, T. (2014) Epigenomic programing: a future way to health? *Microb. Ecol. Health Dis. 25*.
- Shusei Sato, Hideki Hirakawa, Sachiko Isobe, Eigo Fukai, Akiko Watanabe, Midori Kato, Kumiko Kawashima, Chiharu Minami, Akiko Muraki, Naomi Nakazaki, Chika Takahashi, Shinobu Nakayama, Yoshie Kishida, Mitsuyo Kohara, Manabu Yamada, Hisano Tsuruoka, Shigemi Sasamoto, Satoshi Tabata, Tomoyuki Aizu, Atsushi Toyoda, Tadasu Shin-i, Yohei Minakuchi, Yuji Kohara, Asao Fujiyama, Suguru Tsuchimoto, Shin'ichiro Kajiyama, Eri Makigano, Nobuko Ohmido, Nakako Shibagaki, Joyce A. Cartagena, Naoki Wada, Tsutomu Kohinata, Alipour Atefeh, Shota Yuasa, Sachihiro Matsunaga and Kiichi Fukui (2010), Sequence Analysis of the Genome of an Oil-Bear-

ing Tree, *Jatropha curcas* L, DNA Research pp. 1–12
- Silva, G., Legall, J., Xavier, A. V, Rodrigues-pousada, C., Xavier, N. I. O. V, Teixeira, M., & Gall, J. L. E. (2001). Molecular Characterization of *Desulfovibrio* gigas Neelaredoxin , a Protein Involved in Oxygen Detoxification in Anaerobes, *J. Bacteriol.*, 183(15): 4413-20
- Silva, G., Oliveira, S., Gomes, C. M., Pacheco, I., Liu, M. Y., Xavier, A. V., Teixeira, M., LeGall, J., and Rodrigues-Pousada, C. (1999) *Desulfovibrio* gigas neelaredoxin. *Eur. J. Biochem. 259*, 235–243.
- Silva, G., Oliveira, S., Gomes, C. M., Pacheco, I., Liu, M. Y., Xavier, a V, … Rodrigues -pousada, C. (1999). *Desulfovibrio* gigas neelaredoxin. A novel superoxide dismutase integrated in a putative oxygen sensory operon of an anaerobe. *European Journal of Biochemistry / FEBS*, *259*(1-2), 235–43.
- Silva, G., Oliveira, S., LeGall, J., Xavier, a V, & Rodrigues-Pousada, C. (2001). Analysis of the *Desulfovibrio* gigas transcriptional unit containing rubredoxin (rd) and rubredoxin-oxygen oxidoreductase (roo) genes and upstream ORFs. *Biochemical and Biophysical Research Communications*, *280*(2), 491–502.
- Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, Alves LM, Araya JE, Baia GS, Baptista CS, Barros MH, Bonaccorsi ED, Bordin S, Bové JM, Briones MR, Bueno MR, Camargo AA, Camargo LE, Carraro DM, Carrer H, Colauto NB, Colombo C, Costa FF, Costa MC, Costa-Neto CM, Coutinho LL, Cristofani M, Dias-Neto E, Docena C, El-Dorry H, Facincani AP, Ferreira AJ, Ferreira VC, Ferro JA, Fraga JS, França SC, Franco MC, Frohme M, Furlan LR, Garnier M, Goldman GH, Goldman MH, Gomes SL, Gruber A, Ho PL, Hoheisel JD,Junqueira ML, Kemper EL, Kitajima JP, Krieger JE, Kuramae EE, Laigret F, Lambais MR, Leite LC, Lemos EG,Lemos MV, Lopes SA, Lopes CR, Machado JA, Machado MA, Madeira AM, Madeira HM, Marino CL, Marques MV, Martins EA, Martins EM, Matsukuma AY, Menck CF, Miracca EC, Miyaki CY, Monteriro-Vitorello CB, Moon DH, Nagai MA, Nascimento AL, Netto LE, Nhani A Jr, Nobrega FG, Nunes LR, Oliveira MA, de Oliveira MC, de Oliveira RC, Palmieri DA, Paris A, Peixoto BR, Pereira GA, Pereira HA Jr, Pesquero JB, Quaggio RB, Roberto PG, Rodrigues V, de M Rosa AJ, de Rosa VE Jr, de Sá RG, Santelli RV, Sawasaki HE, da Silva AC, da Silva AM,da Silva FR, da Silva WA Jr, da Silveira JF, Silvestri ML, Siqueira WJ, de Souza AA, de Souza AP, Terenzi MF,Truffi D, Tsai SM, Tsuhako MH, Vallada H, Van Sluys MA, Verjovski-Almeida S, Vettore AL, Zago MA, Zatz M,Meidanis J, Setubal JC., 2000, The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis**,** *Nature*. 2000 Jul 13;406(6792):151-9.
- Soares-Castro P, Marques D, Demyanchuk S, Faustino A, Santos PM., 2011, Draft genome sequences of two Pseudomonas aeruginosa clinical isolates with different antibiotic susceptibilities, *J Bacteriol*. 2011 Oct;193(19):5573.
- Soares-Castro, P., Marques, D., Demyanchuk, S., Faustino, a, and Santos, P. M. (2011) Draft genome sequences of two Pseudomonas aeruginosa clinical isolates with different antibiotic susceptibilities. *J. Bacteriol. 193*, 5573.
- Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W., and Shi, L. (2011) Next-generation sequencing and its applications in molecular diagnostics. *Expert Rev. Mol. Diagn. 11*, 333–43.
- Sutton Granger G., Owen White, Mark D. Adams, And Anthony R. Kerlavage, 1995, TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects, *Genome Science and Technology.* 1995, 1(1): 9-19.

- Thoenes, U., Flores, O. L., Neves, a, Devreese, B., Van Beeumen, J. J., Huber, R., Romão, M. J., LeGall, J., Moura, J. J., and Rodrigues-Pousada, C. (1994) Molecular cloning and sequence analysis of the gene of the molybdenum-containing aldehyde oxido-reductase of *Desulfovibrio* gigas. The deduced amino acid sequence shows similarity to xanthine dehydrogenase. *European journal of biochemistry / FEBS 220*, 901–10.
- Thoenes, U., Flores, O. L., Neves, A., Devreese, B., Beeumen, J. J., Huber, R., Romao, LeGall, M. J., Moura J.J.G., Rodrigues-Pousada, C. (1994). Molecular cloning and sequence analysis of the gene of the molybdenum-containing aldehyde oxido-reductase of *Desulfovibrio* gigas. The deduced amino acid sequence shows similarity to xanthine dehydrogenase. *European Journal of Biochemistry*, 220(3), 901–910.
- Turner, N., Barata, B., Bray, R. C., Deistung, J. & Le Gall, J. (1987) The molybdenum iron-sulphur protein from *Desulfovibrio* gigas as a form of aldehyde oxidase, *Biochem.* J. 243, 755-761.

- Vettore, A. L., da Silva, F. R., Kemper, E. L., Souza, G. M., da Silva, A. M., Ferro, M. I. T., Henrique-Silva, F., Giglioti, E. A., Lemos, M. V. F., Coutinho, L. L., Nobrega, M. P., Carrer, H., França, S. C., Bacci Júnior, M., Goldman, M. H. S., Gomes, S. L., Nunes, L. R., Camargo, L. E. A., Siqueira, W. J., Van Sluys, M.-A., Thiemann, O. H., Kuramae, E. E., Santelli, R. V, Marino, C. L., Targon, M. L. P. N., Ferro, J. A., Silveira, H. C. S., Marini, D. C., Lemos, E. G. M., Monteiro-Vitorello, C. B., Tambor, J. H. M., Carraro, D. M., Roberto, P. G., Martins, V. G., Goldman, G. H., de Oliveira, R. C., Truffi, D., Colombo, C. A., Rossi, M., de Araujo, P. G., Sculaccio, S. A., Angella, A., Lima, M. M. A., de Rosa Júnior, V. E., Siviero, F., Coscrato, V. E., Machado, M. A., Grivet, L., Di Mauro, S. M. Z., Nobrega, F. G., Menck, C. F. M., Braga, M. D. V, Telles, G. P., Cara, F. A. A., Pedrosa, G., Meidanis, J., and Arruda, P. (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome research 13*, 2725–35.
- Victor, B. L., Vicente, J. B., Rodrigues, R., Oliveira, S., Rodrigues-Pousada, C., Frazão, C., Gomes, C. M., Teixeira, M., and Soares, C. M. (2003) Docking and electron transfer studies between rubredoxin and rubredoxin:oxygen oxidoreductase. *Journal of biological inorganic chemistry : JBIC : a publication of the Society of Biological Inorganic Chemistry 8*, 475–88.
- Volbeda A, Charon MH, Piras C, Hatchikian EC, Frey M, Fontecilla-Camps JC., 1995, Crystal structure of the nickel-iron hydrogenase from *Desulfovibrio gigas*, *Nature*. 1995 Feb 16;373(6515):580-7.
- Voordouw, G. (1995) The genus *Desulfovibrio*: the centennial. *Appl. Environ. Microbiol. 61*, 2813–2819.
- Wang, Z., Hobson, N., Galindo, L., Zhu, S., Shi, D., McDill, J., Yang, L., Hawkins, S., Neutelings, G., Datla, R., Lambert, G., Galbraith, D. W., Grassa, C. J., Geraldes, A., Cronk, Q. C., Cullis, C., Dash, P. K., Kumar, P. A., Cloutier, S., Sharpe, A. G., Wong, G. K.-S., Wang, J., and Deyholos, M. K. (2012) The genome of flax (Linum usitatissimum) assembled *de novo* from short shotgun sequence reads. *Plant J. 72*, 461–73.
- Withers, M., Wernisch, L., and dos Reis, M. (2006) Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. *RNA 12*, 933–42.
- Zhou, S., Herschleb, J., and Schwartz, D. C. (2007) New High Throughput Technologies for DNA Sequencing and Genomics. *Perspectives in Bioanalysis*, pp 265–300. Elsevier.

ITQB-UNL | Av. da República, 2780-157 Oeiras, Portugal
Tel (+351) 214 469 100 | Fax (+351) 214 411 277

www.itqb.unl.pt